ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA
SCUOLA DI INGEGNERIA E ARCHITETTURA

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA E SCIENZE INFORMATICHE

Semantic web approach for italian graduates' surveys: the AlmaLaurea ontology proposal

Tesi in
Web Semantico

Relatore
Prof.ssa Antonella Carbonaro

Presentata da
Luca Santandrea

Sessione III
Anno Accademico 2016/2017

A Marco, Gina, Rodolfo e Marcello.

A tutti gli amici che mi hanno sostenuto in questo percorso.

Ai colleghi IT, in particolare i developers,

per averne condiviso i momenti più e meno belli.

A Serena, insostituibile compagna di viaggio.

Table of contents

Table of figures

Introduction

Since its creation, the web has certainly had a radical impact on the life style of the people, actually modifying what is the search and the exchange of information. Today indeed it is possible to make use of an enormous amount of knowledge by simply querying a search engine, and this operation is also free from spatial constraints, thanks to the parallel growth of mobile technologies. The current web has furthermore revolutionized many other sectors, like commerce, journalism and telecommunication, fields where currently it holds an increasing and pervasive role.

Anyway, the opportunities made available by this invention conceal an aspect which results as of today only partially solved: the intelligibility and the semantic of the data on the web. If from one side it's true that the web is suitable mainly in a *human friendly* mode, from the other it can't be underestimated its possible use also by the computers, considering also the computational power that they have reached in these years. Moreover, it is known that, thanks to the diffusion of the social networks, of the *Internet of Things* and of the mobile web, the quantity of data generated on the web is surely too high to be entrusted uniquely to a human usage. Not least, it is useful to notice that the current search engines, in fact the principal responsible of the retrieval of the information sparse on the web, found their behaviour on proprietary technologies based mainly on the syntax match of the searched keywords, leading often to the problem of the pertinence of the results. For these reasons it has been introduced the concept of semantic web, which is a web capable to report in a formal manner the meaning of the existing terms, thanks to the definition and the utilisation of ontologies needed to classify the resources, allowing also mechanisms of research and logical inference. These mechanisms consent not only to improve the usage of the contents expressed in the web, thanks to the introduction of a formal structure, but also to demand to the computers reasoning which were not possible before, because of the merely syntactic nature of the web.

Through the semantic web it happens the fill of the so-called *knowledge gap* between human and machine, that is the impossibility for a machine to deduce implicit information inside a context, ability which instead is present in the humans thanks to their bag of knowledge.

The technological standard promoted by the World Wide Web Consortium (W3C) already allow to add a semantic stack to the actual web, and by means of these technologies it is possible the transfer to the *Web of Data*, which is the vision of the web as set of understandable contents expressed in a formal way. These data can be then linked together, realizing the so-called *Linked Open Data,* an interconnected net of freely accessible and usable contents.

The Linked Open Data, together with the motivations previously explained, have caused a growing interest for the semantic web, and the next engagement of firms, research organizations and governments. The latter exactly represent the main promoters of Linked Open Data, purposing to supply a higher administrative transparency and a better support to the proximity between citizens and institutions.

The presented thesis project proposes a referencing ontology for data relative to results of surveys effectuated on Italian graduates, to be used as base for the successive exposing of the data in Linked Open Data format. The final resulting system, based on the AlmaLaurea consortium's annual graduates surveys, consists in a set of OWL ontologies which formally describe the peculiarities of the domain. Together with them, the relative RDF triplestore is provided, dataset which is the result of the process of structuring of the information taken from the AlmaLaurea's questionnaires according to the defined model. The utilization of the dataset is then guaranteed by the support of a specific software which exposes a SPARQL endpoint for the custom knowledge retrieval via the submit of queries. The picture is completed with the release of concrete tools for the structured data usage by the users, in the form of web based data visualization diagrams and forms for the guided creation of the SPARQL queries.

The leading reason of the development of the project can be found in the recent interest of the public administration field on the open data phenomenon; the lack of a semantic web intervention in the field of Italian graduates, and the contemporaneous high presence of public data from AlmaLaurea (which already releases its data in an application driven way) have been exploited to create a definition of an open data version of the survey’s information. This process, in compliance with the AlmaLaurea’s mission of bridging between graduates, firms and institutions, represents an interesting conceptualization of the domain also from an international point of view (as respects the Bologna process directive), and hopes to constitute a referencing help in the promotion of similar initiatives for the continuous growth of the data knowledge spread.

This document recalls the history of the web, the principal technologies related to the semantic web and discusses in depth the motivations and the choices which have been taken for the case study in question. Moreover, several consideration about the usability of the effectuated work and possible growing scenarios are provided.

From human to machine web

The evolution of the web

The birth of the web happened thanks to the intuition of Tim Berners-Lee, researcher at the CERN of Geneva, who in 1989 hypothesized the creation of a more efficient instrument for the exchange of documents and information among the various researchers of the centre. The founding idea of this instrument is the usage of the concept of *hypertext* [1], allowing to documents, univocally identified by a URI, the direct link between them. This operating principle, together with the development of the HTTP protocol, of the HTML language and of the first browser, has contributed to spread the web on a large scale and to take it to general public, making it immediately an interesting source of distributed documents easily reachable. This first implementation of the web consented a *read only* modality, where the users had a passive role with just the obtainment of static documents. The potentialities offered by the web caused therefore also the interest of several companies, that soon tried to impose their own standard, generating the so-called "browser war". For this reason in 1994 it was established [2] the World Wide Web Consortium (W3C), non-governmental organization having the assignment to define and promote the referencing standard technologies of the web.

Years later, thanks to the growing development of applications like blogs and forum, it changed the vision of the web, passing to a version characterized by interactivity with the user and by dynamicity of the pages. This version of the web was renamed Web 2.0, definition that was firstly coined by Tim O'Reilly during the O'Reilly Web 2.0 conference in 2004 [3]. Following this vision, a beginning principle of the web 2.0 was its usage as platform (Web as platform): the possibility to exploit applications and services online moved the process from the desktop environment to the web platform, realizing a substantial modification in the software paradigm and in its distribution, and transforming the desktop computers from elaboration centres to access interfaces to services.

A further prerogative of the web 2.0 was the participation: the possibility of publication, modification and sharing of contents from the single users in the web changed its way of use, which became proactive, and thanks also the creation of software like the *wiki* contributed to the creation of a new common knowledge. This interactivity of the web developed the formation of communties, which are group of persons that actively exchange each other information about determined arguments, and this phenomenon culminated with the creation of social networks, software which brought an enormous impact on the human social relation modalities.

The progressions promoted by the web 2.0 led to an increasing presence of the information on the network, to the point of generate as of today a huge mass of data, surely too big to be managed only by humans. The scientific community has then hypothesized an access to data by the computers, idea that would produce remarkable advantages, such as the creation of a set of services exclusively controlled by the machines, thanks also to the new progresses in the field of artificial intelligence. A further sphere of application could regard the semantic search of information, inducing to an increase of its efficiency and to the solution, in this way, of the problem of mental integration of non pertinent search results.

The efforts spent in this direction are leading to what has been identified as new version of the web, denominated semantic web.

The semantic web

For semantic web it is intended an extension of the web ideated by Tim Berners-Lee, who in an article published on the periodical Scientific American [4] described the opportunity of the passage the Web of Data, namely a web where the information could have a semantic characters such to guarantee the possibility of interpretation and usage by the machines.

This new conception derived from the background error that the web dragged since the dawn, which is the fact that most of the information present in it were thought to be enjoyed only by the humans; in fact, those were unorganized, non structured information not integrated each other, making impossible their reuse from the machines, which unlike the men do not own abstraction capabilities and then are not able to interpret the implicit meaning of contextualized data. In the hypertextual web indeed much of the semantics of data is implicit, deductable for instance from layout, colours or images, elements that a machine is not able to recognize.

It becomes then necessary to restore the initial idea of the web as general space of information, adapt also to the automatic process by the computers. With the semantic web it is possible the passage from the web of documents to the web of data, through the injection in a formal and explicit manner of the meaning of the concepts expressed within the documents, so to make them abstract with respect to the context and adapt for the interpretation by the machines, that do not need no more to base their knowledge uniquely on the syntax. Plus, in this way the data are correctly structured and linked each other, making more efficient the mechanisms of search and integration though their usage from various application.

The semantic web has therefore the objective of giving to the machines the possibility to understand and elaborate the information of the web, letting also the realization of logical reasoning starting from them, going towards the creation of intelligent agents for the support of human activities. This vision has been expressed by Tim Berners-Lee himself in the following quote:

"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize. "

Technological stack

The concrete realization of the semantic web is based on a multi-level architecture, where each level has a different purpose and takes advantage of standard technologies promoted by the W3C. Figure1 provides its simplified representation.

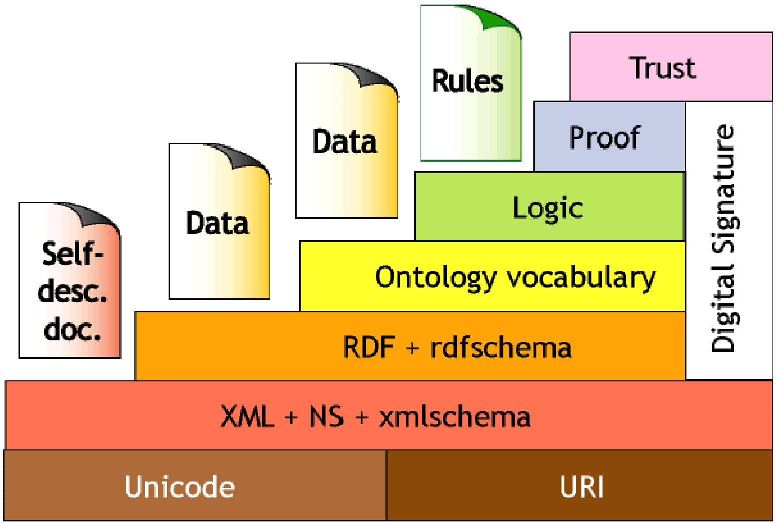


Figure 1 - Semantic Web architecture according to W3C [5]

At the first level there is the actual web, characterized by resources linked each other and identified by an URI (Uniform Resource Identifier), a naming system which allows to assign a univocal name to every resource present on the web [6]. This system is placed side by side at the first level with Unicode, a standardization of the characters encoding independent from platform, language and alphabet [7].

At the upper level is placed the XML (eXtensible Markup Language), an interoperable language though which is emphasized the regulation of contents' syntax [8]. The contents defined by means of XML must indeed respect defined rules which guarantee their syntax correctness (Well formed XML), and these are defined and formalized thanks to the support of formal grammar, like XML Schema [9]. Since a document expressed in XML can be related to many grammars, in order to avoid problems of ambiguity or polysemy, the level is completed with the definition of namespaces, which let the univocal attribution of an identity to elements and attributes used in every XML instance, preventing in this way possible conflicts between different definitions. Thanks to these technologies is therefore possible to realize the syntactical formalization of the metadata for the semantic web.

Since the information presentation modalities are not sufficient to define the meaning, it has been introduced the upper level containing the RDF and RDF Schema standards. The Resource Description Framework (RDF) is a model for the representation of metadata, in which each concept is expressed as a series of triples subject-predicate-object [10]. In each triple (also called statement) the subject is an element equipped with an URI, the object any literal resource (string, date, etc.) or the URI of another element, and the property – which expresses the binary relation between subject and object – has a proper meaning and is in turn defined by an URI. The latter can in fact be collected into vocabularies.

RDF, though the definition of relations between elements, contextualizes the data structures by hierarchical taxonomies and makes it possible the execution of inferential procedures. The schema constructed by RDF is thinkable as a graph where the nodes are the resources and the edges the properties. This graph can then be connected to others, allowing the reutilisation of the formalization of concepts expressed by other sources.

To be able to define the used relations and properties it is employed RDF Schema, a language for the definition of vocabularies [11]. Though it it is also possible to introduce the concepts of graph and hierarchy, applicable both to the objects and to the predicates existing between them.

The following level of the stack is the one defined by the ontologies, key mechanism for the definition of the semantic web, from the moment that it extends the capabilities of RDF schema letting the definition of constraints on the relations defined among the concepts. At this level the standard referencing language is the Web Ontology Language (OWL) [12].

Going up the stack the upper levels are reached, for which there still have not been defined supporting standard technologies. At those levels we find:

- **Logical level**: layer where it happens the passage from the knowledge representation to the application of a logical language and of the relative inference rules, necessaries for the effectuation of reasoning and deduction of new information.
- **Proof level**: layer where are executed the underlying logical rules and are provided explanation on the replies found by automatic agents, needed for their validation.

- **Trust level**: top of the stack, where occurs the verification of the veracity of the obtained information and the trust of the source that makes them available.

RDF and representation formats

The fundamental paradigm for the knowledge representation is implemented into the RDF language. Through it each relation between different objects is described in triples, which therefore allow the definition of statements containing information about a given concept. These statements can even be reified to be exploited as object of new statements (statements about statements) [10].

This mechanism, jointly to the feasibility of usage of standard XML Schema datataypes and of strings equipped with linguistic tags, generates sufficient expressive power to describe in a machine-readable modality metadata on every possible resource currently present on the web. Additional functionalities made at disposal by the language are the possibility of definition of Container (ordered, non ordered or alternatives lists of objects) and of Collections (non extensible lists of objects).

In RDF every single statement can be intended, basing on different points of view, as triple, as sub graph or as a textual code snippet (RDF serialization).The latter modality in particular is due to the usage of RDF/XML as referencing format [13].

Though RDF/XML it is possible to represent each triple exploiting the native syntax of XML: a document having as root a node <rdf:RDF> contains several nodes <rdf:Description> which describe the statements. In particular, the objects of these statements can be literal resources, existing objects (using attribute rdf:about), newly defined resources (using attribute rdf:ID) or blank nodes, anonymous resources useful for instance to define n-ary predicates. The RDF/XML syntax allows eventually using nested descriptions and rules for the abbreviated syntax.

To represent the RDF triples in textual format there are also available different serialization syntaxes; in particular there have been defined:

- **Notation3 (N3)**: a format which lets the serialization of the graphs in a textual modality, resulting in an easier interpretation for the humans [14]
- **Terse RDF Triple Language (Turtle)**: a subset of N3 exclusively dedicated to the simplified serialization of RDF format, which results more compact thanks also to the usage of prefixes [15].
- **N-Triples notation**: subset of Turtle, allows an even more simplified representation of the statements [16].
- **JSON-LD (JavaScript Object Notation for Linked Data)**: it is a specific implementation for the linked data in JSON format, with the purpose to use its existing serialization modalities. It uses a concept of *context* to map properties of JSON objects into an ontology [17].

SPARQL

In parallel to the development of the stack it has been introduced the modality of semantic interrogation of the data. Though the semantic web in fact it is possible to express complex queries, different from those based on keywords typical of the current search engines. In the latter, indeed, it is not possible to express correctly the semantic tie which exists among the different searched terms.

The W3C has promoted the SPARQL standard (Protocol And RDF Query Language), language which permits to research data expressed in RDF format [18]. The SPARQL queries are based on the recognition of patterns over a RDF graph, named path expressions. This is substantially a set of triples expressed in *Turtle* language, that restricts the queried graph returning the information that satisfies it. The pattern triples can contain also variables, bound to RDF terms, used for the print of the results.

Thanks to the possibility of definition of filters, join predicates, sorting and limits on the results, the expressive power of SPARQL is sufficient to execute very complex queries on a RDF dataset. For this reason SPARQL is for RDF what SQL (from which SPARQL is inspired) represents for the relational databases. Finally this language gives the possibility to extend the interrogated knowledge bases thanks to the definition of federated queries: in this way it is possible to use different endpoints contemporaneously in the same query, consenting the opportunity of search between different distributed datasets, used a lot in the Linked Data scenario.

Instruments for the addition of semantic information in the web

As completion of the development of the semantic web technologies, there have been defined some mechanisms for the addition of a semantic layer on the syntactic web, following therefore the idea behind the original philosophy; this result is reached thanks to the usage of specific technologies which allow the insertion of the knowledge directly into the XHTML code of the page, in an embedded mode. In this manner, the addition of metadata related to the syntactic content is oriented to the creation of RDF statements directly starting from the web pages, considerably simplifying the generation of structured data and allowing the reduction of the existing gap between the actual web and the vision of the 3.0 web.

Among the first technologies regarding *semantic markup* there emerge the microformats (μF), HTML code patterns born with the purpose of permitting the addition of semantic information on entities such as persons, events and reviews. Specific microformats regarding various type of information (for instance *hCard* and *hCalendar*) have been developed and promoted by the *microformats.org* community [19]. Their employ results easy since it bases on the usage of HTML attributes *class*, *rel* and *rev*, modality which assures also the maintenance of an excellent readability for the humans. Together with the usage of microformats it can also be associated the employ of the tool GRDDL [20], which starting from them extracts the relative RDF triples.

A second technology for the metadata embedding is RDFa (RDF in HTML attributes) [21]. Differently from the microformats, it is a W3C standard that does not use existing HTML tag attributes, but it defines new specific ones for its purpose. In particular, basing on the fact that the object of the statement is either a literal value or another already defined resource, the used attributes would change. Here follows an example based on the FOAF (Friend Of A Friend) vocabulary:

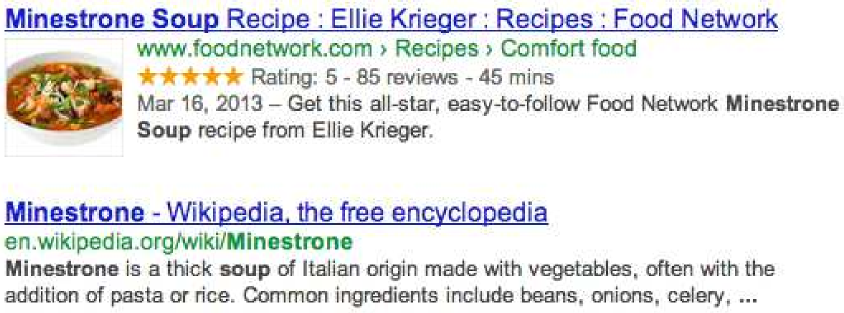
```
<div xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <a about=”http://www.example.com#luca” rel="foaf:homepage"
  href="http://www.lucasantandrea.com/">Luca Santandrea</a>
  <span property="foaf:firstName" content="Luca"/>
  <span property="foaf:lastName" content="Santandrea"/>
```

```
</div>
```

In this example, the defined statements having as object a literal value exploit the attributes *property* and *content*, while the statement having as object another resource with a defined URI uses the attributes *rel* and *href*. RDFa allows therefore to add semantics to the web pages by using a syntax more similar to RDF respect to the microformats, and it has also the advantage of overcoming the limit given by the reduced number of describable typologies of data, being in fact able to refer to all the possible RDF datasets basing on the relative URI.

Another modality for insertion of semantic within the content of web pages is given by the microdata [22], specific originated by the Web Hypertext Application Technology Working Group (WHATWG). In a similar way compared to what RDF does, the microdata exploit a set of vocabularies for the description of resources and their specific properties; in particular, there are usually adopted the vocabularies by Schema.org, web site founded on a common initiative by the main search engines (Google, Bing, Yahoo, Yandex) with the purpose of giving common schemas for the markup of the structured data in the web. Also for microdata there are used specific attributes, like *itemscope* to define the existence of an item, *itemtype* for the definition of a property (which can belong to an external vocabulary) and *itemprop* for the assignment of a given value to the property.

The semantic markup technologies have also brought an immediate feedback even in the search engines: in fact, thanks to instruments like the *Rich Snippet* promoted by Google [23], it has been possible to integrate the structured data directly in the search results, making “richer” the semantic of the obtained contents, allowing the visualization of information such as rating, product prices and article’s authorship. With the diffusion of mobile devices these results have been further enriched, up to be called *Rich Results*, which are search results deriving from structured data showed in dedicated tabs next to the organic results.



Minestrone Soup Recipe : Ellie Krieger : Recipes : Food Network
www.foodnetwork.com › Recipes › Comfort food
★★★★★ Rating: 5 - 85 reviews - 45 mins
Mar 16, 2013 – Get this all-star, easy-to-follow Food Network **Minestrone Soup** recipe from Ellie Krieger.

Minestrone - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Minestrone**
Minestrone is a thick **soup** of Italian origin made with vegetables, often with the addition of pasta or rice. Common ingredients include beans, onions, celery, ...

Figure 2 - Example of Google's organic results with the rich snippets

Finally, the advent of the new HTML5 standard has improved the scenario, promoting the integration of the previously described technologies, defining new accessibility requirements and introducing new structural page tags to let an HTML semantic markup [24].

Challenges of semantic web

The definitions of the standards and of the technologies promoted by W3C aim to face the critical issues that the semantic web has pointed out during its diffusion. In 2008, the W3C Incubator Group published a report [25] where it analyses the challenge of the knowledge representation and of the automatic reasoning, by considering the uncertainty of the information present on the web. To describe the basic behaviour of the uncertain information exchange it has been created an ontology, used also to provide a full coverage about the identification and the classification of the uncertainty typologies. Figure 3 resumes the result of the introduced taxonomy.

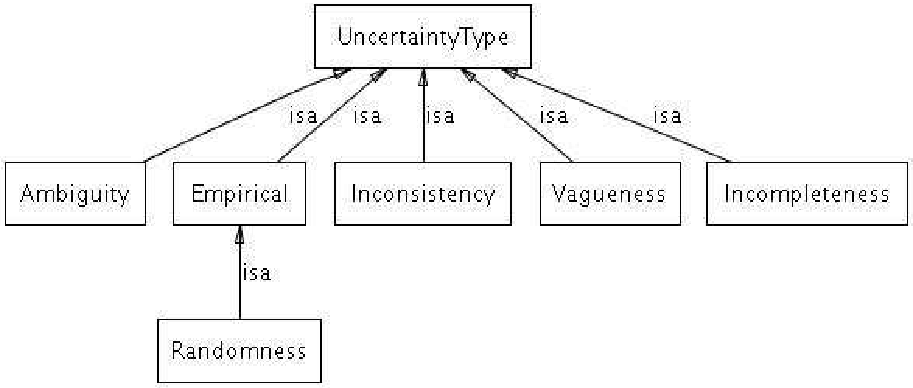


Figure 3 - Semantic web uncertainty typologies [25]

In particular, the recognized typologies concern:

- **Ambiguity**: the references of the terms are not clearly specified, leading to a doubt of their meaning.
- **Empirical values**: the correctness of a term depends on empirical events, so the information is not possessed by the system yet; this event could also have an aleatory nature.
- **Inconsistency**: there are logical contradictions deriving for example from the combination of ontologies coming from different sources.
- **Vagueness**: the expressed concepts are imprecise and are not bound to an exact correspondence in the reality.
- **Incompleteness**: there are necessary further data to define the consistency of the information

Beyond these, it is important to consider aspects like the data uncertainty (intended as lack of precision) and the deception (which is the voluntary provision of erroneous data). Lastly, an important criteria is about the vastness of data: the web in fact contains billions of pages, and it is therefore difficult to correctly classify the resources, since there are also possible semantic duplications.

In the development of the semantic web it is therefore important to provide formally a method for the management of those uncertainties, so that they can be individuated and solved by autonomous agents.

The Linked Open Data project

After the definition of the concepts at the base of the semantic web and of the relative technologies, it has been recognized the possibility of the creation of the *Web Of Data*, a global database of contents accessible by machines, identified in the project called Linked Data. This consists in the connection of information and knowledge exploiting the mechanism of the URI, and allows the link between correlated data where not previously possible.

The linked data are characterized by four simple rules [26]:

1. Usage of URI to identify the elements
2. Usage of URI via HTTP to allow the referencing of the elements
3. Description of the resource in a standard format, for example RDF
4. Inclusion of link to other correlated URI, so to simplify the research of new information.

For Tim Berners-Lee, the Linked Data need the publication of "Raw Data" [27], that are untreated data formally expressed in a way that allows their reuse for other purposes: data belonging to different specific domains can be combined, allowing the discovery of new information and expanding the semantic knowledge on them. In this way it is created a comparison, from the point of view of the data, of what the WWW has represented for the documents.

Open Data

The RDF dataset expressed should also have a nature open to the public, so to guarantee their free availability for everyone. Hence the name *Linked Open Data*.

Through this initiative the purpose is to avoid the "data silos" phenomenon, that is the presence of sources of information confined in private and isolated databases, which can not be reused.

To verify the characteristics of openness of a datum it has been created a chart [26], reported in figure 4:

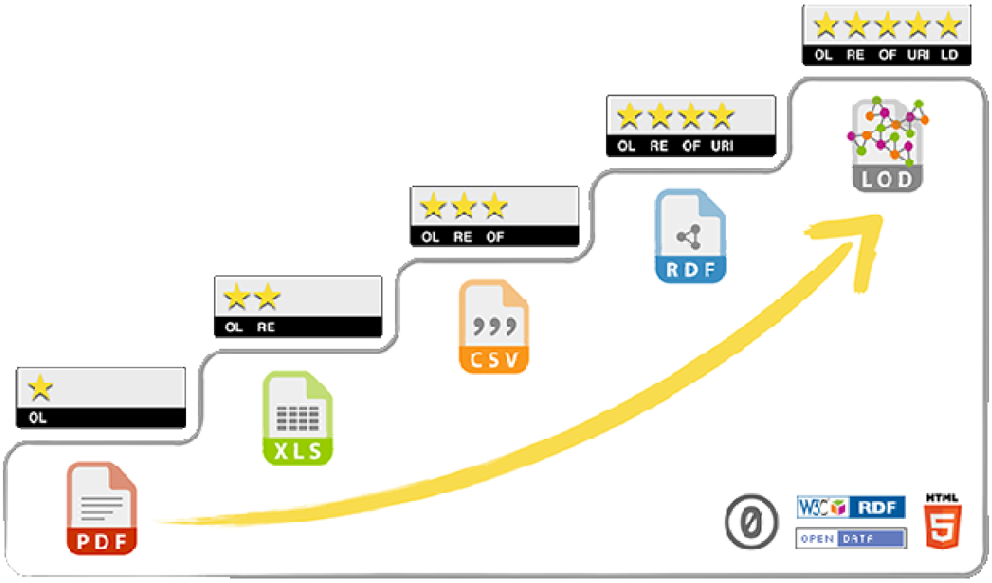


Figure 4 - Increasing hierarchy of the open data typologies

The meanings of the various levels are explained in the following table:

★	Available on the web in whichever format but with an Open license
★★	Available as structured and machine-readable datum (e.g. Microsoft Excel)
★★★	As the previous level, in a non proprietary format (e.g. CSV)
★★★★	As the previous level, in a W3C standard open format (RDF + SPARQL)
★★★★★	As the previous level, including link to connect the data to other open datasets

Among the main producers of open data there are present not only governmental institutions (institutional, administrative, healthcare, etc.) but also users communities. Between these it is worth to mention *OpenStreetMap*, a project aiming at the creation of a dataset with geographical and cartographical information, and *DBpedia*, RDF version derived from Wikipedia in a project published in 2007 by the Free University of Berlin [28].

For their construction, the Linked Open Data start from existing ontologies, like *WordNet, FOAF* and *SKOS*. Their usage, together with the creation of new domain ontologies successively published, has facilitated the parallel generation of *Linked Open Vocabularies* [29], a subset of Linked Open Data panorama regarding the ontologies. This development allows their reutilisation thanks to the import mechanism provided by OWL.

The efforts of W3C have then allowed a growing development of Linked Open Data, leading to the creation of a global RDF graph containing many billions of triples. This graph, named *LOD cloud*, has vertiginously expanded during the years. In figure 5 it is reported a recent representation of it.

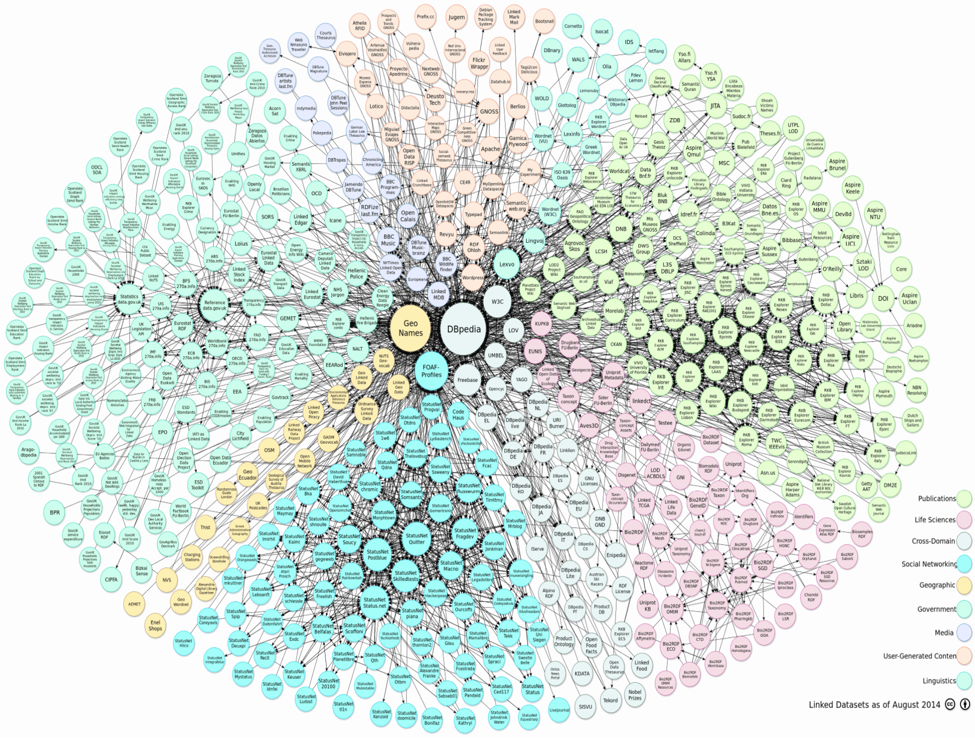


Figure 5 - Linked Open Data Cloud as of August 2014 [30]

SPARQL EndPoint

Thanks to the standard format with which the information are made available in the Linked Open Data it has been possible to develop dedicated SPARQL endpoints so to guarantee the execution of queries. The main implementation technologies for these engines are *OpenLink Virtuoso* and *Apache Jena*.

In order to support further the integration process of the RDF knowledge basis it has been also created the *Pubby* project, having the goal to provide an interface Linked Open Data for the triplestore dataset which are queryable only by means of SPARQL [31].

Some of the most common SPARQL endpoints are reported in the next table:

Bio2RDF	Linked Data for biological sciences
DBPedia	Information parsed by the Wikipediia pages
WikiData	Structured data supporting the creation of Wikipedia pages
Data.gov.uk	UK government data
MusicBrainz	Music database
DrugBank	Database containing information about medicines and active principles
LOD Cloud cache	Endpoint which queries the LOD Cloud

Linked Open Data in Italy

In Italy the Linked Open Data paradigm has spread starting from 2007 with the publication of territory data within the OpenStreetMap project. Later, several independent initiatives have been developed by the users' communities (for instance the website *LinkedOpendata.it*).

The wide range diffusion is reaching in particular thank to the effort of the public administration, in compliance with the *PSI Directive* [32], an European directive of 2003 aimed to regulate the publication and the reuse of the data of the public sector. Although many institutions have promoted initiatives for the publication of open datasets, there are numerous cases that do not provide any supporting SPARQL, and others which not satisfy the RDF format, contributing in this way only to the commitment of the transparency of public administration, but not helping in fact the Linked Open Data project. For these motivations, considering also the fragmentation of the open scenario in the Public Administration, the Italian Government has taken on the responsibility, through the Digital Italian agency, of the publication of the national guidelines for the valorisation of the public sector information [33], of the definition of a license named IODL (Italian Open Data License) and of the creation of a centralized catalogue of the open data of the public administration, in the website *dati.gov.it*. The metadata of this catalogue, which currently collects only a part of the open dataset of the public administration, flow into the *European Data Portal*.

An example of application is given by the ISTAT: in May 2015 it has been published a RDF dataset which exposes data starting from the 2011 census. The website *datiopen.istat.it* makes available a SPARQL endpoint and a GUI to facilitate the users' interaction. Moreover, it is available a REST web service for the integration with external services. The developed dataset uses two different OWL ontologies created ad-hoc: one about the territory data and one about the census data, which have been developed using also references to existing ontologies [34]. Finally, to guarantee the quality of the exposed data it has been used a meta-ontology named PROV-O, which has the purpose to verify the provenance of the exposed data for a better quality assurance.

A possible use of this dataset is given for example from its integration with the Linked Data portal of ISPRA (Italian National System for Environmental Protection) [35]. By means of this link it has been possible to join census data with other regarding indexes of ground consumption, detecting therefore new knowledge about the consumption in determined built areas.

Leaving the governmental sphere, there exist several Italian independent initiatives for the publication of Linked Open Data. Among these there should be mentioned two dataset collections maintained by different users' communities: *DatiOpen.it* and *openDataHub*.

Knowledge representation

To pursue the objective it is necessary to ask ourselves how can a machine interpret the data. The central concepts are the provision of data in a structured way and the existence of inference rules which can be exploited to conduct automatic reasoning. These structures and rules must be also formalized in a standard mode, so to permit their reutilization to everyone. Eventually it is important the flexibility of these structures, to let to all the kinds of data in the web to be represents by means of them.

The scenario of the web leads to the individuation of 3 principal components:

- **Data**: information of any type present into the web
- **Semantic metadata**: information that enrich the content of data, adding to them an interpretable semantic by the machines.
- **Schemas**: formal models that allows to correlate each other metadata through the definition of relations, constraints and class membership rules.

Following this scenario it appears the necessity of the creation of ontologies, to abstract the meaning of the information making it explicit also outside its context.

Ontologies

The term ontology the responsibility derives etymologically from Greek words "ὄντος" and "λόγος", which means argument about being. It concerns a philosophical construct finalized to the discussion and the description of the existence of things, in terms of objects, their relations and relative classifications. In the informatics sphere this name is used to define the formal and explicit representation of a shared conceptualization, according to the definition proposed by Tom Gruber [36].

Rigorous descriptions of objects, concepts and their relations are wrapped and explained by means of ontologies, which have the final goal of expressing formally the knowledge of a given domain. This structured information can be shared and aggregated with other ontologies, for the creation of a greater knowledge domain. In particular, the usage of ontologies can be at the base of a semantic integration among different domains of interest.

The ontology mechanism allows not only the possibility of structuring the data, but also the possibility to make them interoperable and available outside their natural context, adapt for an automatic reasoning. Pidcock depicts how the ontologies can be intended as meta-model useful to describe dataset which models the representation of a domain of interest [37].

Controlled vocabularies, Folksonomies, Taxonomies and Thesauri

The term ontology is usually used within the scientific scope in a univocal manner also for referencing to other knowledge representation modalities, like controlled vocabularies, folksonomies, taxonomies and thesauri. Although not directly expressed into the technological stack promoted by the W3C, these concepts are however part of the web semantic panorama. Over the years, the scientific community has hypothesised various criteria to define their dissimilarities [38][39].

Wong et al. [40] propose a "spectrum" of the different possible ontology typologies, whose representation is reported in figure 6.

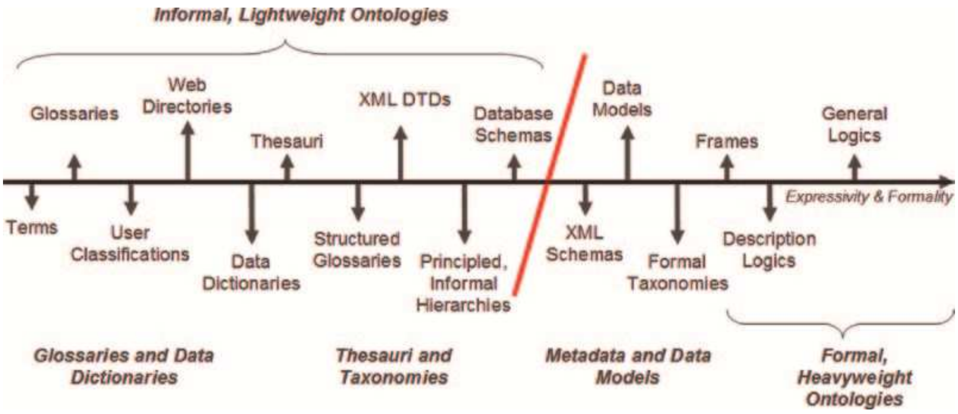


Figure 6 - Ontology spectrum for the semantic web

Controlled vocabularies

Controlled vocabulary means an organized list of terms and sentences initially used to label contents in order to ease their identification after a research. The goal of this classification is to reduce the ambiguity of the terms, associating more names to the same concept [41].

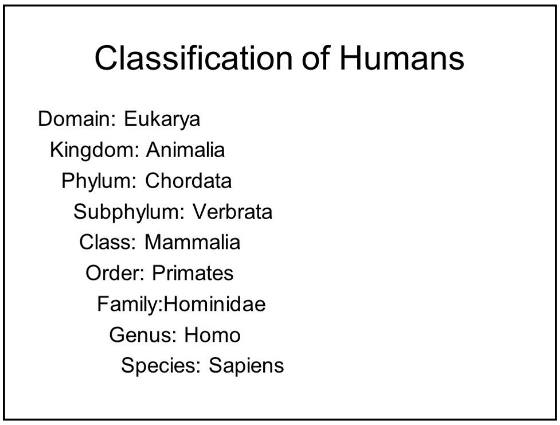
Folksonomies

Folksonomy means a simple list of user-defined keywords to annotate resources on the web. It is a non-formal classification modality, whose diffusion has grown thanks to the social bookmarking mechanisms (of which the most famous example is the website *Delicious*) and to the usage of tag clouds. The simplicity of use, the lack of additional cognitive costs and the extended utilisation by thousands of users in the web decree its importance, regardless the limits due to the lack of structured concepts as hierarchy and synonymy [42].

Taxonomies

A taxonomy is definable as "hierarchical structure to aid the process of classifying information" [43]. The ontologies are frequently reduced to the concept of taxonomy. McGuinnes uses the term "taxonomy" in an equivalent manner respect to the definition of "simple ontology" [44].

A key principle at the base of taxonomy is the utilisation of hierarchical rules among different terms, which are bound each other with "father-son" relations. The referencing example of the concept of taxonomy is the Linnean classification, used to classify the living being in different categories organized in specific hierarchical levels. In this classification it appears clear the father-son relation (generalization of the relation of type "is-a") among elements at different levels.



Classification of Humans

Domain: Eukarya
Kingdom: Animalia
Phylum: Chordata
Subphylum: Verbrata
Class: Mammalia
Order: Primates
Family:Hominidae
Genus: Homo
Species: Sapiens

Figure 7 - Linnean taxonomy for the human being

Thesauri

According to the definition by ISO a thesaurus is “a controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms “.

The purpose is to facilitate the selection of the same term starting from the combination of others, for this reason it is optimized for the usability by the humans [45]. A thesaurus can then be considered as an extension of the taxonomies, where in addition to the relations of hierarchical type, others are made explicit, for instance synonymy and antonymy.

One of the most important thesauri is the Medical subject Headings (MeSH), whose goal is the indexing of terms used in the biomedical sphere of scientific literature.

If the types of relation expressed by thesauri (hierarchical, associative or equivalence) need to be extended, the concept of thesaurus evolves in the most general concept of ontology. The main difference between thesauri and ontologies consists in the fact that the latter base their representation on a formal, logic-based language, whose grammar contains constraints about the usage of the terms and allows successive mechanisms of inference [37].

Classification of ontologies

Within the scope of knowledge engineering there exist several modalities for the classification of ontologies. Guarino [46] proposes a classification based on the level of generality:

- **Top-level ontologies**: describe general concepts such as space, time, subject, object, events, actions, etc. in an independent way with respect to a particular domain of the problem.
- **Domain ontologies**: describe a vocabulary referred to a generic domain (e.g. medicine) specializing the terms provided by the top-level ontology.
- **Task ontologies**: describe a generic process or activity (e.g. selling activity) by specializing the terms given by the top-level ontology.
- **Application ontologies**: describe concepts dependent both from a particular domain and a particular task. These ontologies refer only to a specific application, and in particular the concepts expressed can correspond to roles effectuated by entities during the execution of an activity (e.g. component of a machinery)

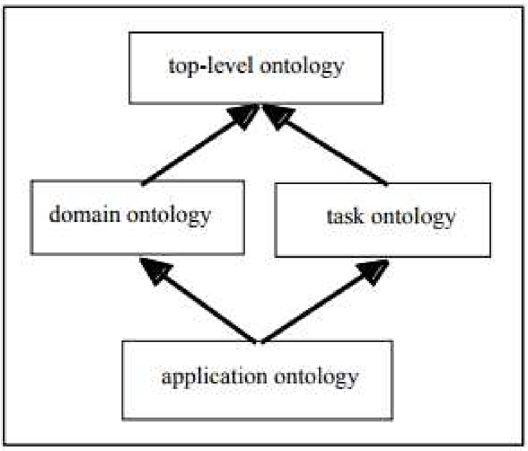


Figure 8 - Ontologies classification according to Guarino [46]

A further modality of classification based on the type of used language for the description of ontologies has been proposed by Slimani [47]. The different individuated categories are:

- **Information ontologies**: composition of diagrams used for the organization of planning ideas of development from different collaborators; it is about not very generic ontologies and very tied to a particular project.
- **Terminological/Lexical ontologies**: ontologies which contain concepts and relations not fully covered by axioms and definitions which guarantee the necessary and sufficient conditions for their usage.
- **Axiomatized/Formal ontologies**: ontologies whose concept and relations have associated axioms. These ontologies require a clear semantic for the used language in order to define the concepts.
- **Software ontologies**: there are ontologies whose goal is to provide conceptual representation focused on data storage and data manipulation, making them adapt to software development activities.

Gomez-Perez and Corcho propose a classification in a lightweight (ontologies which contain concepts, relations and functions) and heavyweight (more complex ontologies that respect to the previous there contain axioms) modality [48].

Eventually, it has been proposed a framework for the construction of ontologies, based on the concepts of semantic dimensions (Language expressivity, granularity, level of structure) and pragmatic dimensions (expected use, automatic reasoning, design methodologies) [49].

Reasoning

Reasoning is intended as a process for the extraction of knowledge starting from an ontology and the related instances. More precisely, this process exploits logical consequences derived by axioms, to infer new facts not explicitly expressed.

The usage of software named *Reasoner* allows the attainment of new implicit information starting from ontological rules and data. The reasoners are also useful to validate an ontology, that is to verify that the rules defined in it do not generate inconsistencies. An example of a Open Source reasoned is HermiT, of which a built-in implementation is present into Protégé, open source software for the design of ontologies, developed by Stanford University [50].

The reasoners which are based on description logic set up their working principles on the fundamental principles of "open world assumption" and "no unique name assumption". Other criteria researched by a classifier are "concept satisfiability, class subsumption, class consistency and instance checking" [51].

OWL

The ontologies are definable in particular thanks to the usage of the Web Ontology Language (OWL); this language has been introduced to face many of the limitations that RDFs set, adding some first order logic constructs. In fact, by using only RDFs (adapt to the definition of simple vocabularies and taxonomies) it was impossible to express concepts or constraints such as the equivalence of classes or the cardinality limitations of some properties.

OWL [12] is a language partially mapped on a description logic, and represents a compromise between the need of a higher expressivity of RDFs and a sufficient decidability necessary for the usage by automatic reasoners. Through it it is then possible to increase the inferences that can be deduced compared to what was possible with RDFs. It is a W3C recommendation which derives from DAML+OIL Ontology Language, deriving in turn respectively from DAML and from OIL, developed from US and European researchers.

Through OWL it is possible to extend the expressivity of RDF and RDF schema, by introducing new constructs to define classes as a function of others, with operators such as union, intersection and complement. Moreover, it is possible to specify mechanisms of equivalence and non equivalence among classes. Eventually, OWL makes it possible to formalize cardinality constraints and advanced properties such as transitive property, functional property and inverse property.

OWL is divided in different syntactical classes:

- **OWL Full**: containing all the constructs of OWL, it is designed for the usage of the syntactic freedom of RDF and compared to the other two versions it limits the expressivity and is undecidable. It can be intended as an extension of RDF finalized at the increase of semantic of the common terms among RDF and OWL.
- **OWL DL**: reduced version that coincides with the maximum subset of the Full version that guarantees decidability, imposing restrictions on the usage of the basic constructs. OWL DL derives from the field of the description logic.
- **OWL Lite**: version further limited which permits the representation of hierarchical classifications and simple constraints, so to improve the efficiency of the reasoners that use it.

Domain ontologies

Within the semantic web scope there have been proposed several categorization taxonomical modalities for what concerns the educational domain. Consequently there are numerous the ontologies present in the panorama, differentiated by content expressivity and specific application scope. An educational ontology recently implemented is TEACH, vocabulary which aims at supporting teachers to link together elements of their teachings. It contains classes and properties at the level of course, module and assignment, in addition to classes to describe students and teachers, allowing then to describe several detailed characteristics of the courses of study. This ontology, developed by the University of Muenster, is projected to be extended with others like FOAF or Dublin Core metadata terms [52].

A further ontology to mention is VIVO, which goal is to describe the academic and research domain [53]. This ontology is directly derived from the homonym web based open source software developed by the Cornell

University, dedicated to the management and the modelation of the activity of scientists and researchers [54]. VIVO integrates external ontologies like SKOS, BIBO and vCard.

Among the most known ontologies in the educational scope it appears AIISO (Academic Institution Internal Structure Ontology), adapt to describe the organizational structure of an academic institution. Its elements are present at different granularity levels, from faculty and institutes to modules and single subjects. AIISO is designed to work in combination with the ontologies AIISO-roles, Participation and FOAF to describe the role of people inside the institutions [55].

It is also possible to make use of the vocabularies offered by schema.org, for a more general structured representation of information concerning the education. The existing schemas can refer to educational organizations and to courses.

The building of domain ontologies in the scope of education is the subject of numerous researches. Ameen et al. explain a process of creation of an ontology about university courses to guide the students in the choice of their career [56]. Furthermore, Dicheva et al propose, after an analysis of the sparsity of the ontologies of the educational domain, the creation of a web platform for their research [57].

For the case of study in object the previously described ontologies can cover only partially the requested representation needs; this is to imply mainly to the fact that they refer to specific domains, and so an external integration can be considered. Due to the particular nature of the domain of interest the thesis project aims to create a specific domain ontology, with possible integrations of external ontologies in the education scope.

Specific ontologies for the education scope

According to the previous overview, it appears clear how the worldwide different typology of organization of educational systems led to the development of different ontologies, each one having different peculiarities related to the specific domain of interest.

Since the application scope of the thesis project regards statistical surveys on courses of study of Italian universities, it is necessary to identify a restriction on educational domain ontologies. In particular, for the universities belonging to the European Union, the subdivision of courses follows the directives defined in the Bologna Process, international reform entered into force in 1999 (currently adopted by 47 countries) having the purpose of creation of an European Higher Education Area (EHEA), characterized by a standardization of the level of the degrees and of the formative credits (ECTS), by a warranty of their equipollence and by the promotion of the international mobility of the students [58].

Demartini et al. have developed a specific ontology named BOWLOGNA adapt to represent the educational domain and consequent to the adoption of the Bologna process. Its creation has followed an incremental process starting form a linguistic lexicon (deriving from the linguistic translation of concept expressed differently in the member countries) which has been later translated in an ontology [59].

The statistics on graduates

The AlmaLaurea surveys

A significant contribute to the graduates statistics in the national sphere is given by the work done by the AlmaLaurea interuniversity consortium. Founded in 1994 after a first project started by the Statistical Observatory of the University of Bologna, the consortium, supported by the Ministry of Education, University and Research, has its main mission in the production of statistical surveys about the situation of the Italian graduates.

The surveys done have a wide representativeness due to the high number of member universities (75 as of the first months of 2018), which guarantees a coverage of more than 90% of Italian graduates. This diffusion made the AlmaLaurea surveys a reference point for the academic community and for the economical and political world. The aspects analysed are divided in two distinct surveys, published annually:

- **Survey on the profile of graduates**: delineates characteristics and performances of the graduates providing a picture of the situation basing on criteria about study condition, satisfaction on study careers and university success (in terms of final mark and regularity of studies). Data derive from questionnaires distributed to students at the end of their course of study and are integrated with administrative documentation coming from the universities.
- **Survey on employment condition of graduates**: monitors the insertion of the graduates in the business world by collecting data deriving from interviews conducted at one, three and five years from the achievement of the degree. Through it it is possible to obtain information about the typology of work done, the average satisfaction, the average retribution and the inherence with the studies.

The data derived from the interviews, effectuated both in telephonic and web modality (CATI and CAWI) and characterized by more than one hundred variables, are publicly available on the AlmaLaurea website and can be consulted in a single modality at different granularity levels. Moreover it is possible to perform comparisons among different collectives, basing on different variables like gender, degree class or degree course.

The AlmaLaurea surveys are presented every year during a dedicated convention, and there are highlighted also observations on specific themes and employment patterns resulting from the interpretation of the data. The high number of effectuated questionnaires (more than 200.000 every year) allows to obtain a significant dataset [60].

Single Annual Report (SUA)

Basing on the data of the AlmaLaurea statistics it is possible to generate reports with information on transparency requirements for each course of study for which it exists the data of at least one graduate within the database. These data concur to the creation of the single annual reports (SUA) for each course of study of each member university of the consortium. The SUA is a management tool useful for the planning, for the realization, for the self evaluation and for the redesign of the course of study, introduced by the law 240/2010. The SUA reports, adapt to express the quality of the courses of study, are published by the National Agency of the Evaluation of The University system and of the Research (ANVUR) and accessible on the platform *UniversiItaly*.

The generation of the reports occurs by selecting a reduced set of indicators starting from the profile and employment condition surveys. Moreover, as for the extended surveys, comparisons and aggregated visualizations with equivalent pre-reform courses are possible.

Other national and european statistical sources

There exist many data sources at national and international level regarding statistics on graduates. An important reference is given by the National Institute of Statistics (ISTAT) which periodically makes available press releases and publishing productions about several arguments including the higher education. In particular there are done sample telephonic surveys every three years in order to monitor the employment condition of the graduates

[61]. A less specific publication is the Italian Statistical Yearbook, a synthetic annual report in whose section dedicated to education and formation are reported information like enrolment, data on degree attainment and about professional placement of graduates [62].

The ISTAT releases its databanks in different possible format for public use, for instance as microdata (collections of elementary data); concerning the scope in object there are available data about the census of graduates and their professional placement [63]. These data, and others deriving from other surveys like the census of the population, are also viewable online on the portal *I.Stat*.

Another statistical data source comes at a ministerial level: the Statistical Office and Studies of the MIUR (USTAT) makes surveys about the world of university and artistic and musical high formation reporting information about the student population, the didactics the institutes and the right to study. These data, together with the national registry of the students (also collected by USTAT), are consultable in an aggregated manner on a dedicated portal, of which a section is reserved to their release in open data format [64].

On the international level it is necessary to mention EUROSTAT, organ of the European Union which processes statistics at community level. Among the published articles regarding the instruction, there are some specific ones like the analysis of the university education’s statistics and the analysis of the graduates’ employment rates in the recent years [65]. In particular, the first reports data like the distribution of graduates basing on sector and gender, starting from data coming jointly from EUROSTAT, from OECD and from the UNESCO statistical office [66]. The EUROSTAT data are published on a dedicated portal and allow the differentiated visualization for each member country. The data collected constitute some useful indicators to monitor the progresses in the persecution of the objectives imposed by the Europe 2020 strategy, political line proposed by the European Council with the purpose of promoting economic growth and sustainability, of which the development of the university education represents a key concept [67].

Linked Open Data for statistics on graduates

Regarding the state of the art of open data availability about the domain in question, the scenario is currently quite fragmented. Taking as reference the portal of the open data of the Italian public administration, a research on the topic “degree” returns only a set of datasets of few specific territorial realities, and therefore does not capture the majority of aspects at a national level. In a similar way, the same non- comprehensiveness problem happens in the European Data Portal, which collects data from the single national sources, and so merges data in a bottom-up modality guaranteeing a standardization thanks to the respect of the principles of the open data paradigm. Despite the difficulties of attainment of a complete picture form the holistic point of view, the open data phenomenon concerning the educational theme is growing and contributes to the creation of a global knowledge which is very important for the future generations [68].

Particularly interesting is the LOIUS project (Linking Italian University Statistics), which proposes the definition of an ontology for the representation of university statistics published by MIUR, by effectuating their exposition in RDFa format with the goal of providing their web-based representation [69].

The AlmaLaurea statistics in the open data

Given the previous scenario, it appears clear how the integration of the AlmaLaurea statistics in the open data scope can give an important contribution to the available information on the status of graduates in our country. The availability of exhaustive information about the graduates’ employment condition perfects those deriving from the ISTAT sample surveys, and enrich them with more specific data thanks to the numerous variables present in the questionnaires. In a similar way, the survey on the profile of the graduates in the open scope further improves the picture, giving exhaustive and reliable information about the quality of the study experience of the graduates, though data collected at the end which also provide a vision of subjective aspects like the personal satisfaction, and that can complete those deriving from the EUROSTAT surveys.

The goal of this thesis project is to structure the AlmaLaurea surveys in order to make them available in an open data modality to complete the *vision* previously described. Due to the high complexity given by the high number of possible dimensions and from the vastness of the database, the analysis and the implementation of the project starts from the SUA reports, which represent the surveys by focusing on a reduced number of variables.

The next challenges of this project regard the extension of the described variables (up to the reach of all of those present into the surveys) and the growing integration with other databases treating the same domain which are present in the open data panorama.

Thesis project

The contexts of application of this thesis project are the semantic web and the linked data. A first analysis of the scenario is made starting from the AlmaLaurea's surveys concerning the graduates' profile and the graduates' employment condition. In this section are analysed the principal steps which lead to the construction of a formal ontology which describes these surveys in a structured way, making possible the expression of the data in a Linked Data compliant format. Moreover, the aim of this thesis s also to propose several graphical tools which help the final user to understand the data and to exploit them to perform particular knowledge extraction actions.

Knowledge representation

Domain analysis

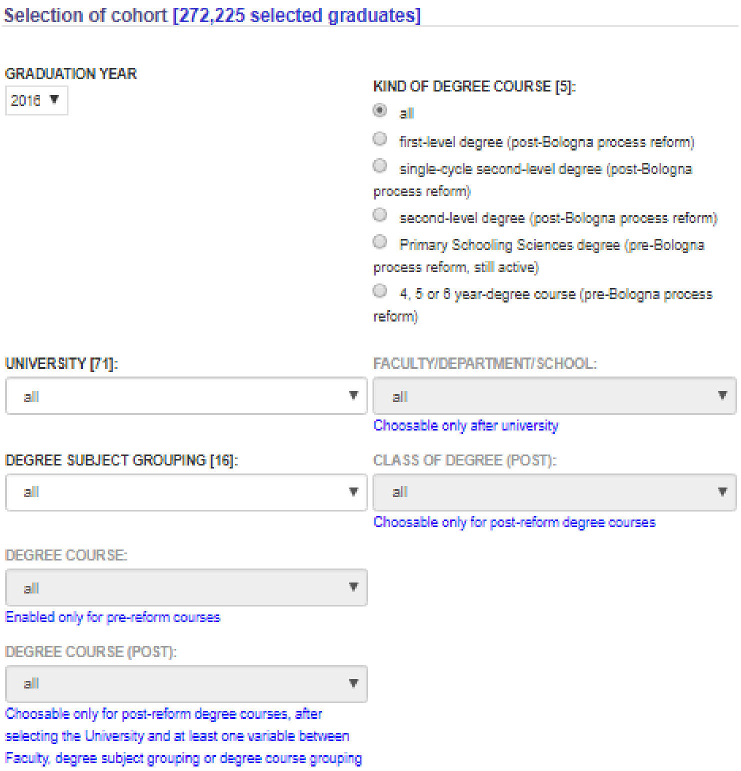
The entry points of the information to be managed are the AlmaLaurea's surveys. These ones are divided into different sections, where each one describes a particular statistical scope and contains several variables which correspond to the questions answered by the students. As an example, the graduate's profile survey is formed by 5 different sections:

- Education and training
- Information on your current course of study
- Evaluation of your current course of study
- Information about your family
- Future intensions and prospects

Each of these contains questions. For instance, within the section 2 ("Information on your current course of study"), there are present questions like "How many of the classes did you attend on a regular basis?" (named R105) or like "Did you study abroad?" (named ESTERO).

Data coming from the submitted interviews pass a data cleaning phase, and therefore are validated throughout the appliance of statistical rules and grouped in defined indexes (process explained in a methodological notes document [70])

The collected data are available on the AlmaLaurea website and are currently queryable by using different dedicated user interfaces, following fixed hierarchical criteria selections. For instance, starting at the university level, it is possible to restrict the research by selecting a particular faculty to extract the data of the surveys of a limited data subset. The subset selection is completed with the possibility to compare the results over different dimensions like the full University datum, the gender division or the different year of enrolment.



Selection of cohort [272,225 selected graduates]

GRADUATION YEAR
2016 ▼

KIND OF DEGREE COURSE [5]:
● all
○ first-level degree (post-Bologna process reform)
○ single-cycle second-level degree (post-Bologna process reform)
○ second-level degree (post-Bologna process reform)
○ Primary Schooling Sciences degree (pre-Bologna process reform, still active)
○ 4, 5 or 6 year-degree course (pre-Bologna process reform)

UNIVERSITY [71]:
all ▼

FACULTY/DEPARTMENT/SCHOOL:
all ▼
Choosable only after university

DEGREE SUBJECT GROUPING [16]:
all ▼

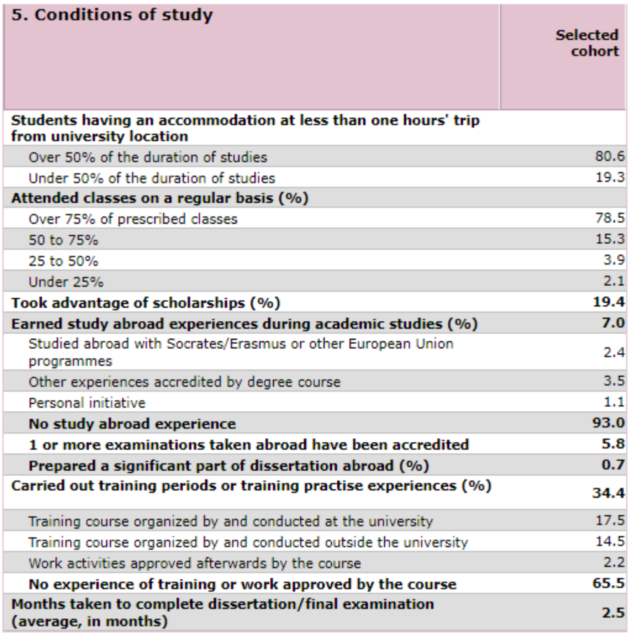
CLASS OF DEGREE (POST):
all ▼
Choosable only for post-reform degree courses

DEGREE COURSE:
all ▼
Enabled only for pre-reform courses

DEGREE COURSE (POST):
all ▼
Choosable only for post-reform degree courses, after selecting the University and at least one variable between Faculty, degree subject grouping or degree course grouping

Figure 9 - AlmaLaurea's graduates profile query form

The performed parameterization of the search form leads to a result page containing the aggregated values regarding the values of different variables of the survey, limited to the selected cohort. The results can also be exported in a CSV format for further analysis.



5. Conditions of study	Selected cohort
Students having an accommodation at less than one hours' trip from university location	
Over 50% of the duration of studies	80.6
Under 50% of the duration of studies	19.3
Attended classes on a regular basis (%)	
Over 75% of prescribed classes	78.5
50 to 75%	15.3
25 to 50%	3.9
Under 25%	2.1
Took advantage of scholarships (%)	**19.4**
Earned study abroad experiences during academic studies (%)	**7.0**
Studied abroad with Socrates/Erasmus or other European Union programmes	2.4
Other experiences accredited by degree course	3.5
Personal initiative	1.1
No study abroad experience	**93.0**
1 or more examinations taken abroad have been accredited	**5.8**
Prepared a significant part of dissertation abroad (%)	**0.7**
Carried out training periods or training practise experiences (%)	**34.4**
Training course organized by and conducted at the university	17.5
Training course organized by and conducted outside the university	14.5
Work activities approved afterwards by the course	2.2
No experience of training or work approved by the course	**65.5**
Months taken to complete dissertation/final examination (average, in months)	**2.5**

Figure 10 - Results of the search in graduates profile survey (fifth section)

From a technical point of view, the available data is deriving from different tables stored in a *Microsoft SQL Server* database. In particular, the data is grouped in a TSQL view where different dimensions are defined to represent each record according to the purposes. Data source includes table of the graduates' registry, the table with the information about the related courses and the records of each specific survey performed every year.

This view effectively is the result of the deployment of a given fact table, whose main characteristics are described in the schema in figure 11:

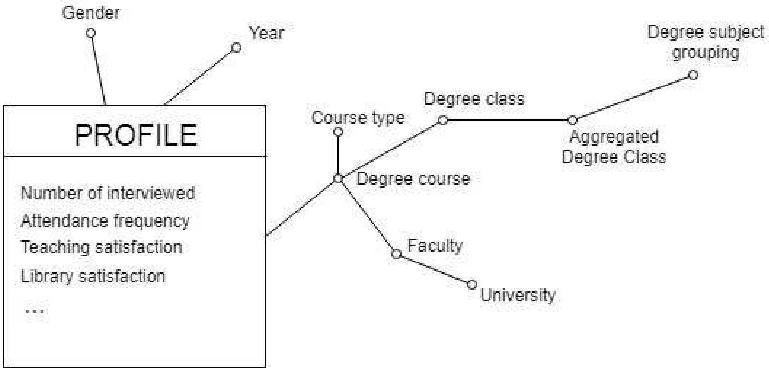


Figure 11 - Simplified dimensional fact model for describing the profile fact

The modelled fact table is principally described by the course dimension, whose hierarchy lets a roll-up analysis into less granular level. The same paths are therefore available in the graphical reports.

From a wider point of view, the profile fact table reported is a data mart derived from the AlmaLaurea's Data Warehouse, a central "information heritage" which combines information coming from different sources such as member universities' administrative data, website access logs or companies' graduates curricula search statistics, in addition to the data elaborated from the surveys. A more detailed description of the AlmaLaurea's DW [71] and technical implementation details [72] have been presented at the 12th European University Information System (EUNIS) conference.

Despite the knowledge representation power offered by the visualization tools, the resulting data about the surveys suffer of lack of freedom of navigation: in fact, the logic used to obtain the data is driven by defined paths strongly dependent on the tools themselves. Furthermore, the results are presented in an aggregated way, by returning counting, sums or percentages of the variables mapped in a defined scenario.
Following these considerations, it appears clear how the information are stuck within the representation software, and therefore it is considerable the issue of facing the *information silos* problem.

The idea of giving a semantic structure to these information has the purpose to overcome the limits of the previous model. The decision to make the data more expressive by adopting the semantic web technologies aims to get the rid of the application dependency, letting the final user capable to freely obtain information in an interoperable way without the constraint imposed by a software. In this way the data becomes usable in different contexts, for instance as basis of mashup applications, gaining also the possibility to enhance their value by combining them with external sources.

In order to better capture the fundamental concepts about the surveys, and therefore to be able to correctly describe the main dimensions available in a full manner, the decision is to focus the efforts on a reduced part of the whole set of possible statistic variables offered in the surveys. This is a starting point to concentrate the work on a limited scope to generally describe the structure in a formal way, avoiding to represent the whole characteristics of the survey (work that goes beyond the purposes of this thesis) but describing in a general way the principal aspects of the AlmaLaurea surveys.

The adopted set of variables to be considered for the project development has been identified in the SUA. This sheet, whose purposes have already been discussed in the introduction part, maintains the same general structure of the full surveys while reporting only a subset of the whole variables of the questionnaires. In particular, it delineates 10 different questions for the profile survey and 6 different questions about the employment condition survey. For this reason, the SUA sheet has been chosen as the target of the modelling phase.

Towards a semantic scenario

The idea of the project is to redefine the information of the SUA in a semantic way, by adding metadata and following a precise structure. For doing this, a workflow has been defined. Starting from the reified cube of each fact table connected to a specific survey, the data are extracted and transformed in RDF format following a RDF/XML syntax, according to the rules of a defined ontology. After that, the generated triple store file is uploaded on a triple store server, which has the task of interpret the data and to provide an endpoint for the query, performed via SPARQL. The structured data can also be used to release a graphical user interface able to better explain certain queried facts.

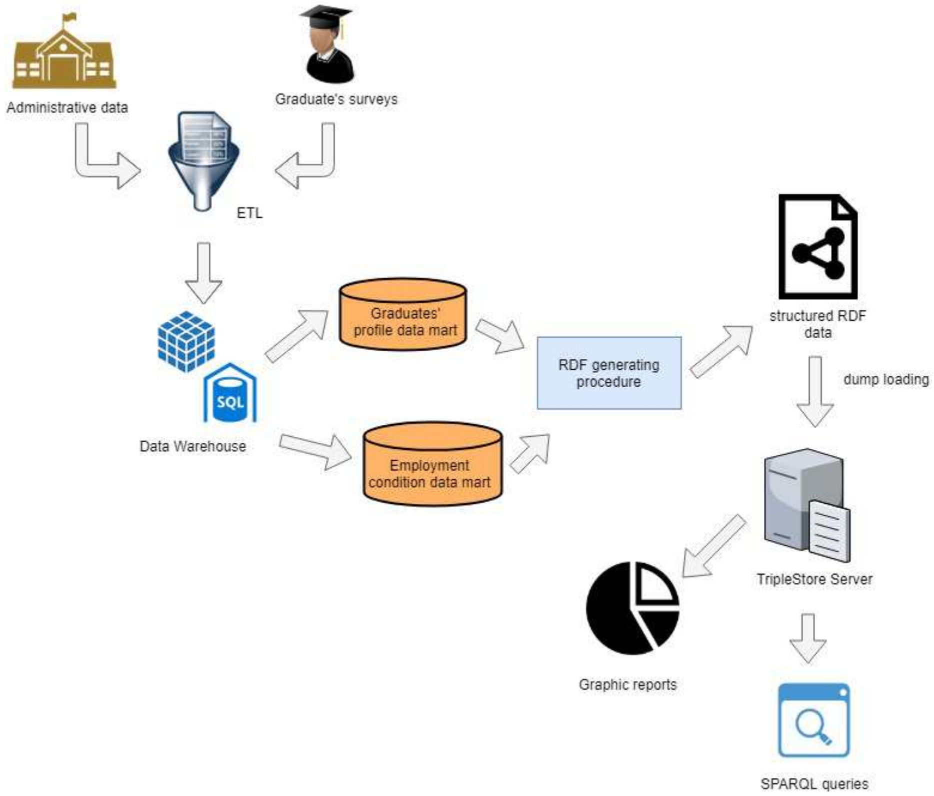


Figure 12 - Generic architecture for the thesis project

A fundamental part of the hypothesized architecture is represented by the ontology, which must express in a formal and unambiguous manner all the aspects that have to be stated into the RDF. As not all the traits reported in the SUA sheet have a covering reference ontology, an *ad hoc* ontology has to be developed, supposing also the import of external third-part ontologies validated and adapt to represent specific dimensions. As an example, the variable about percentage of class attendance of the students of a course, referring to a question on the profile questionnaire, is too specific and needs to be defined in a new ontology. Conversely, the information about a degree course can be described by using well defined ontologies already available in the literature. Some examples of candidates ontologies to be used have been reported in the first chapter, and they regard specific conceptualizations of concepts in the educational domain (like course or institution description).

Variables mapping

A first consideration to be done is about the variables of the SUA. In fact, these actually are present as fields of the database tables which represent each survey data mart. Therefore, the information stated by each column is currently implicit and so not understandable by a non-human user. To overcome this flaw, a variable mapping is necessary. By doing this, the desired variables are listed and enriched with the addition of metadata which describe them in an unambiguous way. This process is then made to formally define all the predicates which will be used in the next RDF definitions, deriving them directly from the variables of the surveys (i.e. the columns of the fact tables).

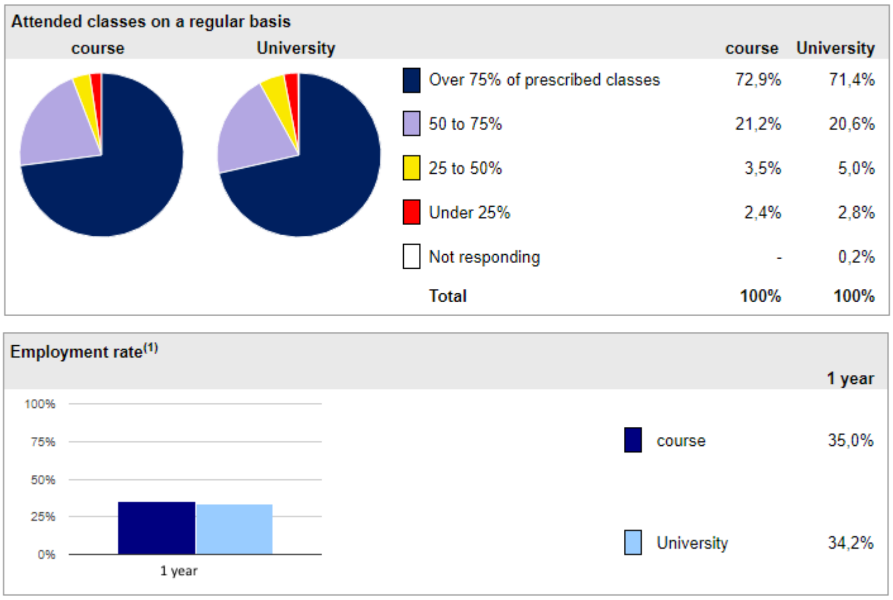
For instance, the following table reports an extract of how the variables of the SUA profile sheet (referring the columns of the database table) are mapped into a new semantic format. In addition to some general variables, there are reported the dimensions of the possible values of the first question identified as R105 (Attended classes on a regular basis).

DB Column name	RDF property name	Comment	OWL Type	Range
Codicione	PROFILO_CORSO	Degree course	ObjectProperty	Corso
Classe	PROFILO_CLASSEDILAUREA	Class of degree	ObjectProperty	ClasseDiLaurea
Anno	ANNO_INDAGINE	Survey year	DatatypeProperty	xsd:gYear
Numlau	NUMLAU_RECENTI	Number of graduates (since 2011)	DatatypeProperty	xsd:integer
interv_1	NUM_INTERVISTATI_RECENTI	Number of interviewed graduates (since 2011)	DatatypeProperty	xsd:integer
numlau_tutti	NUMLAU	Number of graduates (total)	DatatypeProperty	xsd:integer
interv_1_tutti	NUM_INTERVISTATI	Number of interviewed graduates (total)	DatatypeProperty	xsd:integer
regol_0	NUMLAU_IN_CORSO	Number of graduated within prescribed time	DatatypeProperty	xsd:integer
r105_1	r105_1	Less than 25%	DatatypeProperty	xsd:integer
r105_2	r105_2	25 – 50%	DatatypeProperty	xsd:integer
r105_4	r105_4	50 – 75%	DatatypeProperty	xsd:integer
r105_5	r105_5	More than 75%	DatatypeProperty	xsd:integer
r105_0	r105_0	Not answering	DatatypeProperty	xsd:integer

The mapping process has the purpose to list the possible predicates in order to define a corresponding representation in semantic format. Specifically, the individuated predicates have been renamed (in a human friendly manner) and for each one a textual description has been provided. Depending on the type of property, whether linking individuals to either individuals or data values, an OWL specific type has been assigned. These types, defined into the Owl reference specification, are subclass of RDF class *rdf:Property*. It can be noted that the current mapping may refer not only to data values, but also to object ones. For this reason, new kinds of object individuals must be defined. Consequently, these objects' peculiarities will be described by the definition of specific classes, which are introduced in the next section.
The last column reported in the mapping table concerns the range of the property, that is the type of resource which will be the target object of the rdf triple having as predicate the property taken in exam. For the DataTypeProperty predicates, the type inserted are described by using the XML Schema defined data types. Every range is related to its *rdf:Property* by the property *rdfs:range*.

The same mapping activity has also been done over the employment condition survey, enriching the scenario with the addition of metadata on the variables of the questionnaire. In this case, differently from the previous, the subject of the properties is the concept of "Graduate's employment condition".
The two tables have been taken as the main knowledge source for the definition of the domain ontology.

A final analysis has also been done on the visual report already implemented for the visualization of the SUA data. In this case, the study is done focusing on the use of the data, in order to capture the work of the visualization tool so that also this aspect of the data is formalized, letting the description of both structure and behaviour of the information. Even in this case there are two different surveys whose data are represented in web-based charts. These reports file, written in PHP language, build their logic on retrieving the data from the database and extracting the variables ready to be exploited by a graphic library.



Attended classes on a regular basis

	course	University
■ Over 75% of prescribed classes	72,9%	71,4%
🟪 50 to 75%	21,2%	20,6%
🟨 25 to 50%	3,5%	5,0%
🟥 Under 25%	2,4%	2,8%
□ Not responding	-	0,2%
Total	100%	100%

Employment rate[1]

	1 year
■ course	35,0%
🟦 University	34,2%

Figure 13 - SUA report visualization example for profile and employment questions

The query extracted variables are actually the same used for starting the mapping process. However, by observing the query implementation it is possible to recognize how the survey data carry many domain restrictions, which knowledge is hidden in the boundaries of the software. The constraints individuated concerns three different main criteria:

1. **Privacy issues**: regardless the fact that the data are shown in an aggregated form, for cohort with very low cardinality it is necessary to envisage a different visualization. In particular, the system conceals results for sets having less than 5 individuals.
2. **Source integration**: the source table, already normalized and cleaned, may have a lack of consistency if cross-sectioned among different sources. In the AlmaLaurea system, this problem is mainly caused by the integration of the 2014 integration of Vulcano-Stella consortium survey data [73].
3. **Different year versions**: Similarly to the previous point, the surveys performed may vary over different years. A wide temporal range could not satisfy the presence of all the variables, as they could have been added in newer questionnaire versions.

These criteria, together with other more specific (applied basing on the portion of data extracted) have been utilized also during the writing of the procedures needed for the triple store generation.

For this thesis project the collective selected refers on a 3 year basis, retrieving the data up to 2015 survey. This decision follows the reflection on constraint listed before, and has been taken to maintain the highest possible level of data consistency. Besides, this reduction helps the performances, as by just considering these years the resulting graph counts more than 3 million triples.

A first domain ontology proposal

Successive to the mapping of the variables, the construction of the ontology needs the definition of the principal subjects which refer to the defined predicates. In this way, together with the previous analysis of the scenario, the main classes are individuated; Its names and characteristics are listed next:

- **Profilo**: class representing the profile statistics. Its instances are the subjects of the triples regarding the profile survey. This class has many datatype properties, regarding the possible values of the related questions in the questionnaire.
- **Occupazione**: class representing the employment condition statistics. Similarly to the profile class, its instances are defined by datatype properties concerning the survey variables.
- **Corso**: class which represents a degree course. Due to the generality of the concept, many of its characteristics can be expressed by using properties defined in already defined ontologies
- **Ateneo**: class for the definition of University institutions. As the course class can be defined by other ontologies.
- **ClasseDiLaurea**: represents the degree class. This concept, specific to the Italian educational system, serves as a grouping method for similar degree courses, letting an horizontal division with regards to the university hierarchical system. Likewise the course and the university classes, this object is a dimension of the profile data mart.

After the definition of the main classes, there have been hypothesised the basic relations among them. In particular, the link between the different classes has been defined as *owl:ObjectProperties* as they link individuals to other individuals. A brief graphic representation of these connection is visible in figure 14.

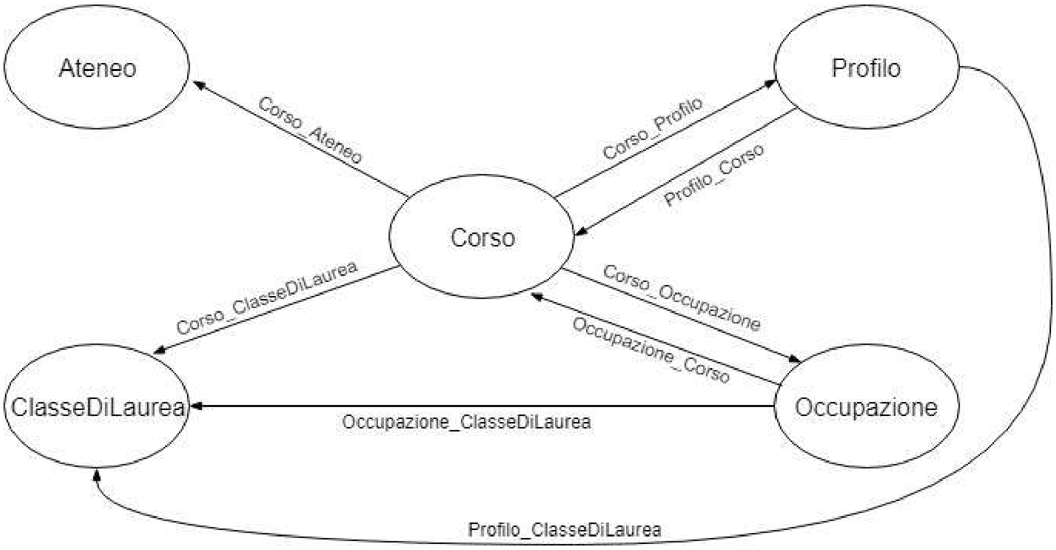


Figure 14 - Classes and relative relations

For both profile and employment objects there have been defined properties for linking the related course. Moreover, for these properties there are also present the inverse relations, so that the navigation of the graph can start from the course. This is actually the same entry point of the SUA graphical reports.

The starting ontology has been defined by merging the rules pointed out with the variable mappings and the others deriving from the link of the main classes of the schema. Further modifications are analyzed in the next sections, when changes of the scenario will be reflected on the conceptualization of the model.

An example of data structured according to the new ontology is reported as follows.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE rdf:RDF [
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
]>
<rdf:RDF

xmlns="http://www.almalaurea.it/opendata/ontologies/almalaurea
#"
```

```
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

  <!-- Example of Course -->
  <rdf:Description rdf:ID="0370106200800008">
      <rdf:type rdf:resource="#Corso" />
      <Corso_Profilo
rdf:resource="0370106200800008_2008_2016"/>
      <Corso_Ateneo rdf:resource="70003"/>
      <Corso_ClasseDiLaurea rdf:resource="2008"/>
      <CorsoCodicione>0370106200800008</CorsoCodicione>
      <CorsoDescrizione>Corso di Laurea in Ingegneria e
Scienze Informatiche</CorsoDescrizione>
      <CorsoSedi>CESENA</CorsoSedi>
  </rdf:Description>

   <!-- Example of University -->
   <rdf:Description rdf:ID="70003">
      <rdf:type rdf:resource="#Ateneo" />
      <AteneoCodice>70003</AteneoCodice>
      <AteneoDescrizione>Università degli Studi di
BOLOGNA</AteneoDescrizione>
      <AteneoSitoWeb>http://www.unibo.it/</AteneoSitoWeb>
  </rdf:Description>

  <!—Example of Degree Class -->
  <rdf:Description rdf:ID="2008">
      <rdf:type rdf:resource="#ClasseDiLaurea" />
      <ClasseDiLaureaCodice>2008</ClasseDiLaureaCodice>
      <ClasseDiLaureaCodiceMin>L-8</ClasseDiLaureaCodiceMin>
      <ClasseDiLaureaTipo>LT</ClasseDiLaureaTipo>
      <ClasseDiLaureaDescrizione>Laurea in Ingegneria
dell'informazione</ClasseDiLaureaDescrizione>
  </rdf:Description>

  <!-- Example Profile -->
  <rdf:Description rdf:ID="0370106200800008_2008_2016">
      <rdf:type rdf:resource="#ProfiloPerCorso" />
      <Profilo_Corso rdf:resource="0370106200800008"/>
      <Profilo_ClasseDiLaurea rdf:resource="2008"/>
      <ANNO_INDAGINE>2016</ANNO_INDAGINE>
      <NUMLAU>87</NUMLAU>
      <NUM_INTERVISTATI>85</NUM_INTERVISTATI>
      <R105_0>0</R105_0>
      <R105_1>2</R105_1>
      <R105_2>3</R105_2>
      <R105_4>18</R105_4>
      <R105_5>62</R105_5>
      <R105_RISPONDE>85</R105_RISPONDE>

      (…. many others datatype properties)

  </rdf:Description>

  <!-- Example Employment condition -->
  <rdf:Description rdf:ID="0370106200800008_2008_2016">
      <rdf:type rdf:resource="#OccupazionePerCorso" />
      <Occupazione_Corso rdf:resource="0370106200800008"/>
      <Occupazione_ClasseDiLaurea rdf:resource="2008"/>
      <ANNO_INDAGINE>2016</ANNO_INDAGINE>
      <NUMLAU>87</NUMLAU>
      <INTERV_1_LAV>85</INTERV_1_LAV>

      (…. many others datatype properties)

  </rdf:Description>

</rdf:RDF>
```

Technical support

The development of the semantic version of the SUA data has been made with the support of several instruments, whose contribute has been fundamental for the achievement of the goal. The semantic web software panorama is quite large, and many different solutions are currently proposed both from private corporations and from open source institutions. For the ongoing project the utilised tools are about the managing of the triplestore server, the provision of a SPARQL endpoint and the building of the ontology. In this section it is reported an overview of the chosen software products and their use for the project purposes.

Protégé

Albeit the first ontology has been constructed from scratch directly with a code editor, a possible growth of the ontology can be difficult to manage. For this reason the use of the Protégé software has been adopted. This tool is an open source editor [50] which makes it easier to define ontologies by the presence of tabs for the editing of characteristics like hierarchy relations, annotations or advanced OWL constructs such as inverse, functional and transitive properties. Thanks to the presence of an internal reasoned, this software could be used also to perform reasoning processes, inferring knowledge starting from the given ontology. The use of Protégé has been very important in this project to simplify the redesigning of the ontology due to the introduction of the comparison among collectives, explained later.

TSQL stored procedures

Driven by the defined ontology, the building of the RDF triple store has been made by exploiting the previously cited SQL queries of the reports. These

have been modified as the extracted values have been encapsulated inside triples expressed in RDF/XML format. The bulk execution of these queries led to the creation of a RDF file ready to be uploaded on a RDF engine.
As each query is bounded to a given course (due to the WHERE selection predicate), an improvement has been done by taking the logic inside a TSQL stored procedure. The latter creates a cursor which extracts all the degree codes of a given university, and then iterates the creation of the triple over all the set. In this way the performed procedure depicts a more structured and general way to build the triples.
The role of the stored procedure is central in the project thesis, as it represents the RDF generating procedure of the architecture as shown in figure 12. Specific details on the different construction approaches are discussed in next sections.

Apache Jena Fuseki

Apache Jena is an open source Java framework for the building of semantic web and linked data applications. Its environment provides APIs for the construction of RDF graphs and the serialization of the triples in various formats. The support of RDFS and OWL guarantees the improvement of the semantic definitions, and is maintained also in built-in reasoners.
The choice of this tool has been made due to the full features offered and for the presence of *Fuseki,* a built-in SPARQL server [74]. The latter (previously named J*oseki*) offers an accessible HTTP endpoint which is exploited in the project for the management of the triplestore.

The usage of Fuseki consists in uploading the generated RDF files on the server; the engine parses and interprets the file defining an abstract model of the graph. The possibility to perform SPARQL queries helps to double check the cardinality of the dataset, the values of the uploaded files and the consistency of the data structuring. Apart from these usages, via SPARQL is possible to define queries to extract specific data from the dataset, going beyond the limits imposed by the previous reporting tools. Here follows an example of SPARQL query used for retrieving, within the profile survey, the average percentage value of the replies “more than 75%” for the question “Attended class on a regular basis” for the top 100 courses.

```
PREFIX alma:
```

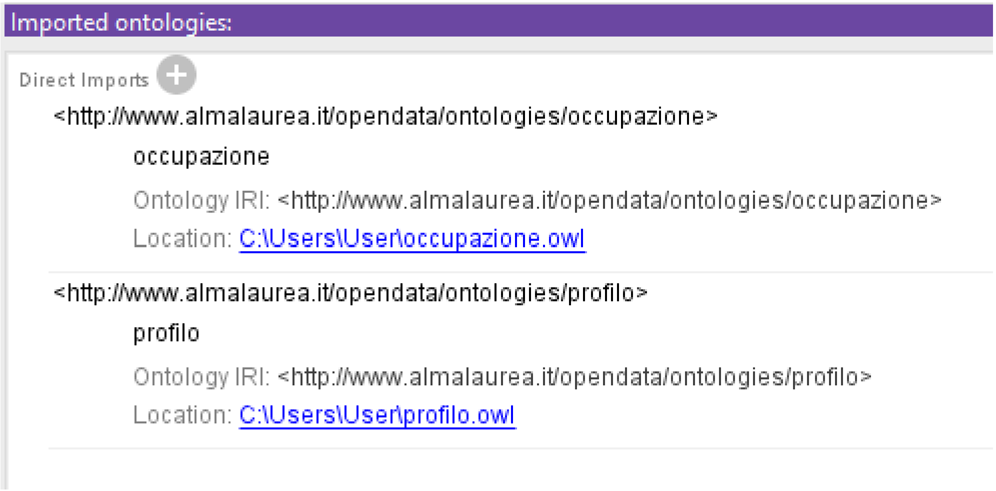
```
<http://www.almalaurea.it/opendata/ontologies/almalaurea#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (AVG(?media) as ?mediatotale) WHERE{
  SELECT
(xsd:float(xsd:integer(?giudizio1)/xsd:integer(?giudizio2)) as
?media)
  WHERE {
    ?profilo rdf:type alma:Profilo.
    ?profilo alma:R105_5 ?giudizio1.
    ?profilo alma:NUM_INTERVISTATI_MENORECENTI ?giudizio2.
  }
  LIMIT 100
}
```

Ontology clarification

In order to better define the different scopes of the treated domain, the existing ontology has been split into three separated ones. This decision aims at clarifying the boundaries among the main concepts individuated, easing the understanding and the management. The different scopes individuated are the following:

- **Profilo**: ontology which contains the definition of the profile class (Profilo), its datatype properties and its object properties.
- **Occupazione**: ontology which contains the definition of the employment condition class (Occupazione), its datatype properties and its object properties
- **Default**: ontology not bound to a specific survey, serving the definition of the remaining classes (Corso, Ateneo, ClasseDiLaurea) and their related properties.

The generation of three different ontologies defines, for each of them, the related namespace, which can be used for the declaration in the header of other ontologies.
With this modifications the object properties ranges refer to objects defined in diverse ontologies, so it has been necessary to apply the import of the ontology via the *owl:imports* statement, in order to use the classes defined elsewhere together with all the rest of the connected semantic (e.g. class/property hierarchy definitions).



<http://www.almalaurea.it/opendata/ontologies/occupazione>
occupazione
Ontology IRI: <http://www.almalaurea.it/opendata/ontologies/occupazione>
Location: C:\Users\User\occupazione.owl

<http://www.almalaurea.it/opendata/ontologies/profilo>
profilo
Ontology IRI: <http://www.almalaurea.it/opendata/ontologies/profilo>
Location: C:\Users\User\profilo.owl

Figure 15 - Import of ontologies into the default ontology using Protégé

The new structure of the ontologies has then been used to correctly describe the data deriving from the SUA reports. Specifically, the reports capture four different surveys, grouped in the two kinds of visualization previously described (profile and employment condition). The surveys whose data have been represented in a semantic format are the following (taken as example the 2017 published surveys):

- **Profile**: measuring the performances about the 2016 newly graduates.
- **Employment condition (one year)**: monitoring the employment condition one year after the graduation (taken in 2016 on 2015 graduates)
- **Employment condition (three years)**: monitoring the employment condition three years after the graduation (taken in 2016 on 2013 graduates)
- **Employment condition (five years)**: monitoring the employment condition five years after the graduation (taken in 2016 on 2011 graduates)

According to the differences between the three employment surveys (e.g. possible diversity of representation or meaning of variables) the related stored procedures for the generation of RDF have been updated. This editing have taken into account also the fact that, with the new domain application, an object of class “Occupazione” could refer to either a 1-year, 3-year or 5-year employment condition survey’s instance.

Contemporaneously to the refinement of the ontologies and of the RDF dataset, a first analysis has been done for the provision of a visual interface for reporting a summarization of the data in graphical chart. A detailed analysis of the process is present in the next chapter.

Substantial modifications to let comparison of collectives

Having the purpose to reproduce the same expressivity of the existing SUA reports, an aspect to be introduced is about the comparison between the single course values and the values coming from less granular levels: university and degree class.

Collective cardinality

A first adversity encountered regards the different cardinality of the dataset to be compared. Actually, a simple sum among all the degree courses values of a given cohort can’t represent a good comparison set. This happens because the counting of this sum and the actual number of graduates of the other cohort (university or degree class) does not return the same value. This mismatch is due to the privacy restrictions described before; in this way, all the courses whose surveys have been filled by less than 5 graduates (threshold value) are not present into the triplestore; thus, the sum of all the single values of a given cohort can be slightly different from the real total count, generating an inconsistency in the data interpretation. To solve this issue, a different organization of the classes has been introduced. The main idea is to provide a dedicated link between the instance of the survey objects (profile, employment) and the comparison cohorts (university, degree class), bypassing the link though the single degree instance. After this approach, the following classes have been generated:

- **Profilo** defines three subclasses:
    - **ProfiloDiAteneo** (Profile for university)
    - **ProfiloDiClasse** (Profile for degree class)

    - **ProfiloDiCorso** (Profile for a single degree course), representing the legacy behaviour described before (degree course centric view)
- **Occupazione** defines three subclasses:
    - **OccupazioneDiAteneo** (Employment condition for university)
    - **OccupazioneDiClasse** (Employment condition for degree class)
    - **OccupazioneDiCorso** (Employment condition for a single degree course), representing the legacy behaviour described before (degree course centric view)

Aggregated values

Another aspect to be considered regards the possible request of visualization of the aggregated values. This option consists in the addition of the values of a previous version of a selected course. Basing on the organization of the degrees ruled by the M.D. 270/04 [75], the recent courses adopt a different naming and organization schema with respect to the previous reform, the D.M. 509/99 [76]. Since the aggregated version of a course integrates the values of the previously related course, the aggregated version of a class instead sums also the values of all the degrees whose next version is a degree of the analysed class.

The analysis of this scenario led to the generation of other classes:

- **ProfiloDiClasse** generates its subclass **ProfiloAggregatoDiClasse** (Aggregated profile for degree class)
- **ProfiloDiCorso** generates its subclass **ProfiloAggregatoDiCorso** (Aggregated profile for degree course)
- **OccupazioneDiClasse** generates its subclass **OccupazioneAggregataDiClasse** (Aggregated employment condition for degree class)
- **OccupazioneDiCorso** generates its subclass **OccupazioneAggregataDiCorso** (Aggregated employment condition for degree course)

For the university cohort the aggregated value is implicit, as the values reported already take into account all the possible courses of a university (and so also the old version ones).

Collective comparison

Going deeper into the analysis of the collectives, according to the behaviour of the SUA reports, it is possible to notate other particularities which further discriminate the comparison sets. A first characteristic is the fact that both individuated cohorts can be divided again basing on “the kind of degree” variable. Following the Bologna process guidelines, the degrees have three possible disjoint levels assigned:

- First level: all possible kinds of bachelor
- Second level: master degrees and single-cycle master degree
- Third level: doctorate programs

The importance of this additional specification is given by the fact that the actual behaviour of the SUA report is to compare data from a single course with other coming from a cohort showing only values of data having the same kind of degree of the single one. This refinement makes the comparison action even more clear and precise. This new scenario conducts to the introduction of other more specific classes, defined as subclasses of those representing the values of the university and the class in both the profile and the employment surveys. The newly defined classes capture a subset of its parent classes limited to the three degree levels previously defined.
As an example, for the profile value of a university, there have been defined the three subclasses **ProfiloDiAteneoL**, **ProfiloDiAteneoLS** and **ProfiloDiAteneoLSE**, representing respectively the first, second and third level of the degree as stated in the Bologna Process.

A particular definition of these subclasses has been done for the classes representing the degree class profile (ProfiloDiClasse) and the degree class employment (OccupazioneDiClasse). As these can also contain aggregated values (option valued with the definition of ProfiloAggregatoDiClasse and OccupazioneAggregataDiClasse classes), other classes have been defined for the specific subset of the latter to maintain high level of expressivity (e.g. for OccupazioneAggregataDiClasse the three subclasses

OccupazioneAggregataDiClasseL, **OccupazioneAggregataDiClasseLS** and **OccupazioneAggregataDiClasseLSE** have been generated).

A more expressive ontology model

The result of the application of the previous considerations has led to the creation of more specific classes, whose employ better represents the domain, giving the possibility of performing an easier knowledge extraction and comparison among different collectives. Consequently to these modifications, also the starting model has changed. In figure 16 it is possible to see the new classes' relations concerning the profile survey domain.

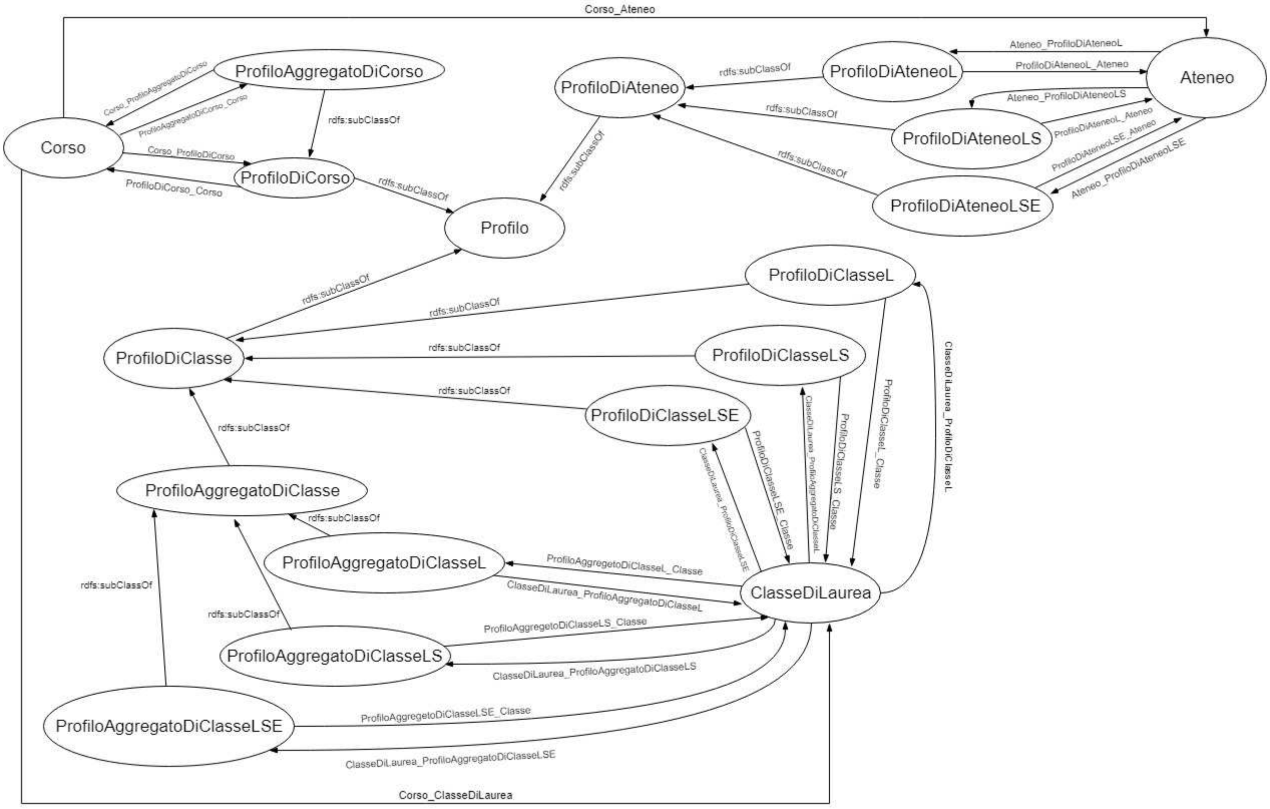
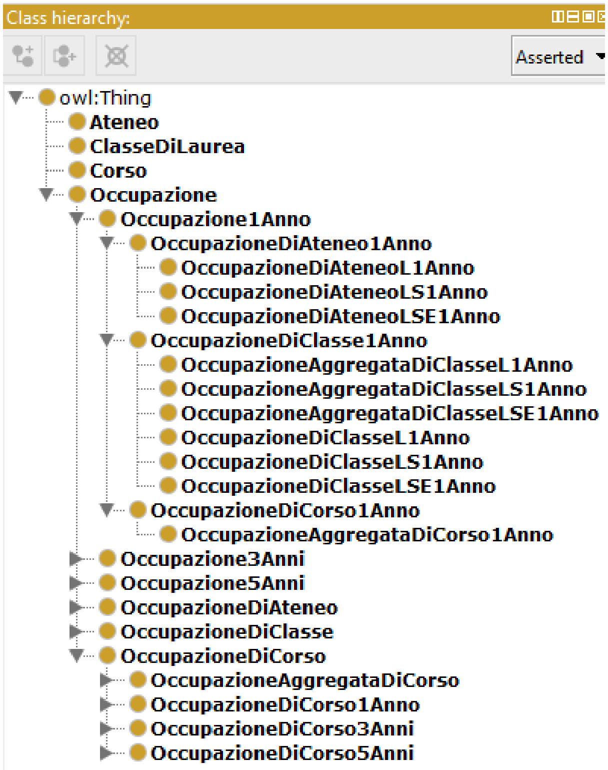


Figure 16 - Updated classes and relations schema for the profile survey

Together with the growth of the number of classes, also a naming convention for the IDs of the generated instances has been proposed, to simplify the possible human reading. Each ID is created as a combination of other codes of related objects. In the following table are reported the naming convention adopted for the main classes related to the profile survey.

Class	ID pattern
ProfiloDiCorso	Degree Code _ Degree class code _ Year
ProfiloAggregatoDiCorso	Degree Code _ Degree class code _ Year_"AGGR"
ProfiloDiAteneo	University code _ Year _ Kind of degree
ProfiloDiClasse	Degree class code _ Kind of degree _ Year
ProfiloaggregatiDiClasse	Degree class code _ Kind of degree _ Year _ "AGGR"

From the point of view of the employment survey, similar measures have been applied for the update of the ontology. Contrastively from the profile survey, however, the employment scenario is characterized also by the time variable. Indeed, the three different types of questionnaires refer to three separated kind of "Occupazione" objects: for example looking for the same degree course, in the same survey year would return values for the survey at one, three and five years. For this motivation, the present classes have been further refined with the creation of specific ones bounded to a given survey year. This action has strongly incremented the number of the available classes. To better construct the class and property hierarchies, and to verify their correctness, the aid of Protégé has been very important. In figure 17 it is reported the final class schema for the "Occupazione" ontology.



```
owl:Thing
  Ateneo
  ClasseDiLaurea
  Corso
  Occupazione
    Occupazione1Anno
      OccupazioneDiAteneo1Anno
        OccupazioneDiAteneoL1Anno
        OccupazioneDiAteneoLS1Anno
        OccupazioneDiAteneoLSE1Anno
      OccupazioneDiClasse1Anno
        OccupazioneAggregataDiClasseL1Anno
        OccupazioneAggregataDiClasseLS1Anno
        OccupazioneAggregataDiClasseLSE1Anno
        OccupazioneDiClasseL1Anno
        OccupazioneDiClasseLS1Anno
        OccupazioneDiClasseLSE1Anno
      OccupazioneDiCorso1Anno
        OccupazioneAggregataDiCorso1Anno
    Occupazione3Anni
    Occupazione5Anni
    OccupazioneDiAteneo
    OccupazioneDiClasse
    OccupazioneDiCorso
      OccupazioneAggregataDiCorso
      OccupazioneDiCorso1Anno
      OccupazioneDiCorso3Anni
      OccupazioneDiCorso5Anni
```

Figure 17 - Occupazione ontology's classes after the change of scenario

A final note regards the definition of the hierarchies into the employment ontology. In this case it has been observed that the newly generated classes can derive from different parent classes. As an example the class regarding the university's first level degrees values of the 1-year employment survey (class OccupazioneDiAteneoL1Anno) can be determined both from the university first level degrees values (OccupazioneDiAteneoL) and from the university 1-year values (OccupazioneDiAteneo1Anno). Thanks to the support of the RDFS language, a multiple superclasses option has been adopted.

Concomitantly with the refactoring of the classes, the definition of the RDF creation via the stored procedures has changed. In particular, it has been defined a single stored procedure for each leaf-level generated class. This led to a notable number of separated scripts, all of these having similar traits. In order to better organize the work, a factorization has been applied.

Final implementation

The factorization process has been done to create a reduced set of queries to be launched directly on the database in order to return the triplestore file in RDF/XML format. The basic idea behind this process is to use TSQL cursors to iterate throughout the survey tables, in order to create a set of triples for every possible class present into the questionnaire data. By doing this, two different points of view have been individuated for the slicing of the data marts. The first one regards the degree courses and university values. The generated stored procedures in this case adopt nested cursors. The algorithm's pseudo code for the generation is reported as follows:

```
-define the requested year
-query and extract all the possible universities from the surveys
-foreach university
     -get all the degree codes of the given university
     -foreach degree
         -generate the profile/employment values
         -generated the aggregated profile/employment values
         -generate the information of the course, and its object properties to bind the profile/employment object previously defined
     -end
     -print the profile/employment values (for the university) at the first, second and third degree level
     -generate the information of the university, and its object properties to bind the profile/employment previously defined
-end
```

The other entry point is about the degree class. In this case, the slicing is made in a cross-university mode, so nested iteration is not necessary. The pseudo code of this second algorithm is the following:

```
-define the requested year
-query and extract all the possible degree classes from the surveys
-foreach degree class
    -generate the profile/employment values (for the class) at the first, second and third degree level
    -generate the aggregated profile/employment values (for the class) at the first, second and third degree level
     -generate the information of the degree class, and its object properties to bind the profile/employment previously
```

```
defined
-end
```

These processes let the creation of all the RDF triples of the surveys published in a given year. This modular work has led to the final definition of 12 stored procedures, reported in the next table.

Stored procedure name	Output
Occupazione_aggregata_classe_1A	Aggregated values for class employment at 1 year, related class information
Occupazione_aggregata_classe_3A	Aggregated values for class employment at 3 years, related class information
Occupazione_aggregata_classe_5A	Aggregated values for class employment at 5 years, related class information
Occupazione_classe_1A	Values for class employment at 1 year, related class information
Occupazione_classe_3A	Values for class employment at 3 years, related class information
Occupazione_classe_5A	Values for class employment at 5 years, related class information
Occupazione_corso_1A	Values for degree employment at 1 year, related course information, Values for university employment at 1 year, related university information
Occupazione_corso_3A	Values for degree employment at 3 years, related course information, Values for university employment at 3 years, related university information
Occupazione_corso_5A	Values for degree employment at 5 years, related course information, Values for university employment at 5 years, related university information
Profilo_corso	Values for degree profile at 1/3/5 years, related course information, Aggregated values for degree profile at 1/3/5 years, related course information, Values for university employment at 1/3/5 years related university information
Profilo_aggregato_classe	Aggregated values for degree class profile at 1/3/5 years, related class information
Profilo_classe	Values for degree class profile at 1/3/5 years, related class information

As completion of the picture, more knowledge has been added to the automatic extractions explained. This static data have been created manually and formalized in RDF/XML format so to be capable to be uploaded in the JENA's Fuseki server. For each of these proposals, specific properties have been added to the classes.
The newly generated data are about:

- **Previous / next version of a course**: this information, taken from a decoding table which stored all the equipollent degree courses, has been comfortable as fallback while searching information in several older surveys (e.g. 5 years surveys from 2014 – data of 2011)
- **Inter-class courses**: due to the fact that an Italian university course can be assigned to multiple classes, this kind of information is useful in order to avoid inconsistencies in data interpretation (especially if taking into consideration the class point of view)
- **University regions**: basing on the *GeoNames* ontology, each university has been marked with the corresponding region. This particular information has become useful within the construction of the GUI for the visualization of the data.
- **University dimension**: similar to the previous, this information asserts which is the dimension of a given institution. This data could be useful in fact it better specifies the representativeness of a particular extraction.

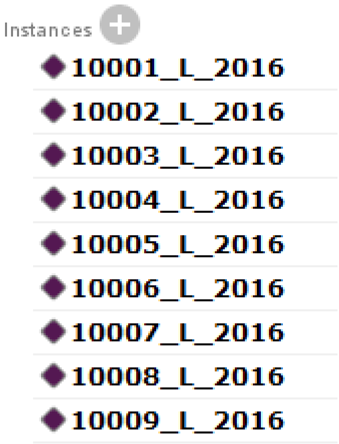
The final purpose of these additional improvements is to introduce new aspects which could be interesting for a further analysis of the data, for the identification of particular patterns or the more precise comparison of the performances.

Reasoning

After the final implementation of the ontology, the last issue to face regards the reasoning. The main idea is to exploit the reasoner tool of Protégé to validate the ontology and to verify the knowledge discovered by the software through the inferential mechanisms. With the reasoning process, three different objectives are reached: the checking o the consistency (with the explanation of unintended relationship between objects), the automatic classification of instances in classes and the equivalence of classes or properties.

First evaluation

A first evaluation performed with the HermiT reasoner essentially confirms the lack of inconsistencies of the generated ontology. This fact can be ascribed to the relative absence of particular properties or generalization within the ontology. As described in this chapter, the final ontology represents a hierarchical order of classes basing on the granularity level. The expressed hierarchy, in this case, is quite standardized as simple direct *rdfs:subClassOf* and *rdfs:subPropertyOf* properties have been used. Moreover, each defined subclass has specific sub properties derived from the ones of the related superclass. The consistency of the ontologies has also been confirmed by the import of RDF instances, whose construction respected the ontology rules and therefore not led to errors.


Instances

◆10001_L_2016
◆10002_L_2016
◆10003_L_2016
◆10004_L_2016
◆10005_L_2016
◆10006_L_2016
◆10007_L_2016
◆10008_L_2016
◆10009_L_2016

Figure 18 - Import of ProfiloDiClasseL instances in Protégé

Usage of OWL advanced constructs

In order to guarantee a full use of the defined model, the ontology has been improved with the addition of advanced constructs allowed by the OWL language. In this way the reasoning process can consider also other kind of relations, increasing the possible paths to analyze to infer new implicit knowledge.

Inverse properties

The first operation regards the definition of the inverse relation on the existing properties. By defining this relation it is possible to perform the subject-object navigation in both directions, increasing the knowledge power. For instance, the property which links an instance of the profile survey for a degree with the related instance of the degree (ProfiloDiCorso_Corso) is explicitly declared as the inverse of the property which links an instance of the degree with its instance in the profile survey (Corso_ProfiloDiCorso)

Disjoint classes

The definition of the disjunction between the classes aims at partially overcome the open world assumption which characterizes the semantic web. The declaration of the disjoint classes avoids ambiguous multiple class affection of instances, whose separation is specific and clear in the existing ontology. This restriction is useful to help the reasoner to individuate unwanted behaviors of the data.

Functional properties

Among the defined data and object properties no restrictions have been defined concerning the cardinality. In order to limit the possible number of definition of a statement for a given property the functional relation has been added. By doing this, only one possible object can be defined for each object, and so this restriction helps the reasoner to point out possible inconsistencies. For what concerns the object properties, they have been defined only for the properties which have an instance of a survey (profile or employment) as domain. The reason for this choice is the fact that, for the cited cases, the inverse functional property is not valid; for example, a given course can be related to more instances of the profile survey for a degree (e.g. for different years), whereas the opposite is invalid.

Disjoint properties

In a similar way for what done with the classes, the disjoint relation has been made explicit also on the object properties. This mechanism strengthen the concept of membership of a given instance to a class, limiting not only the possible *rdfs:subClassOf* property values but also the existence of object properties to only the allowed (not disjoint) ones.

Ontology verification

After the addition of the explained properties, the reasoner has been launched again. A first thing to notice regards the inference of hierarchical relations among classes and properties. Through the work of the reasoner there are pointed out inferred relations, which increase the knowledge base. Figure 19 reports an example of this process.

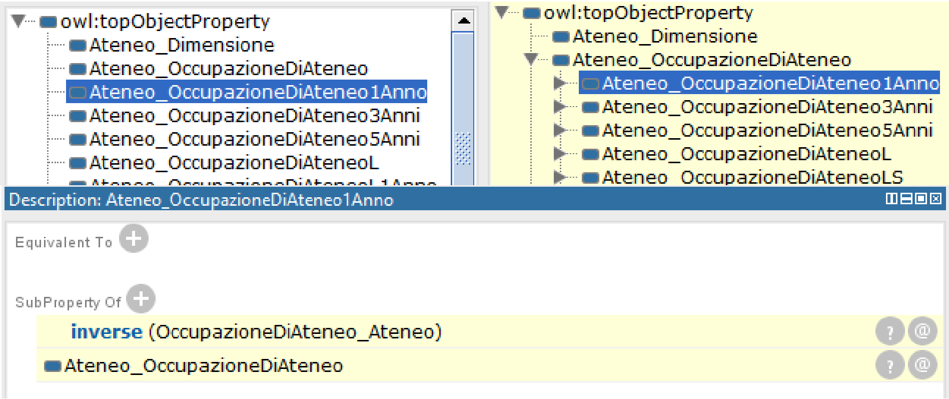

owl:topObjectProperty
Ateneo_Dimensione
Ateneo_OccupazioneDiAteneo
Ateneo_OccupazioneDiAteneo1Anno
Ateneo_OccupazioneDiAteneo3Anni
Ateneo_OccupazioneDiAteneo5Anni
Ateneo_OccupazioneDiAteneoL
owl:topObjectProperty
Ateneo_Dimensione
Ateneo_OccupazioneDiAteneo
Ateneo_OccupazioneDiAteneo1Anno
Ateneo_OccupazioneDiAteneo3Anni
Ateneo_OccupazioneDiAteneo5Anni
Ateneo_OccupazioneDiAteneoL
Ateneo_OccupazioneDiAteneoLS
Description: Ateneo_OccupazioneDiAteneo1Anno
Equivalent To
SubProperty Of
inverse (OccupazioneDiAteneo_Ateneo)
Ateneo_OccupazioneDiAteneo

Figure 19 - Inferred subclass hierarchy of the Ateneo_OccupazioneDiAteneo property (highlighted in yellow)

The usage of the reasoner after the definition of the advanced constructs is also useful to verify the consistency of the objects and of the instances. As the previous evaluation has not pointed out any error, a test has been effectuated by inserting an object intentionally wrong. For instance, the definition of a class (named *WrongClass*) as subclass of both Profilo and Occupazione reports an error, as the two classes have been declared as disjoint. Figure 20 reports the inconsistency highlight in Protégé.

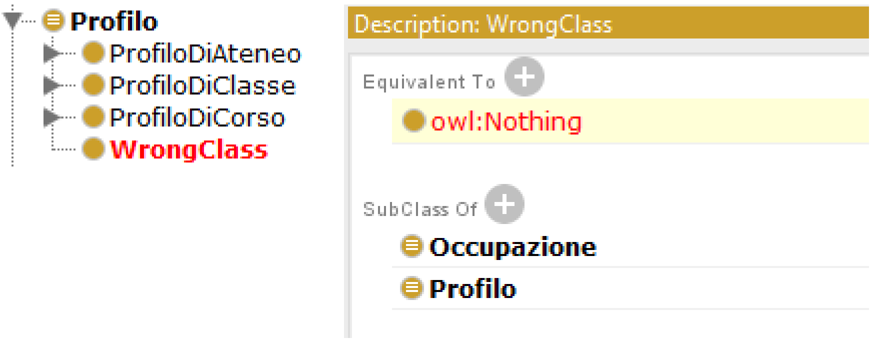


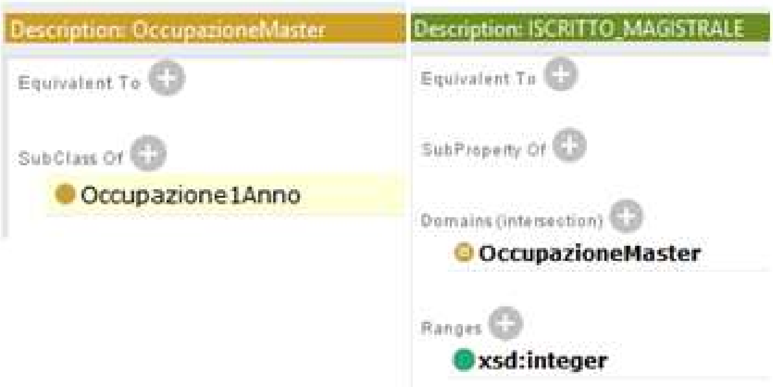
Figure 20 - Inconsistency found in the ontology

Union of classes

Another possibility made available by OWL is the definition of class as union of classes. In this way, a class is defined starting from the characteristics of others, exploiting the OR Boolean combinator.
As an example, in the employment condition survey a particular class has been individuated, which is the class of all the employment survey entities having the data property regarding graduates which are enrolled to a master degree course (named ISCRITTO_MAGISTRALE). As this property can be linked only to subjects who refer to first-level courses, the data property domain has been changed from the Occupazione class to the new defined class, called *OccupazioneMaster*, which is the generated from the union of the classes where the data property ISCRITTO_MAGISTRALE can be satisfied:

- OccupazioneAggregataDiClasse1Anno
- OccupazioneDiAteneo1Anno
- OccupazioneDiClasse1Anno
- OccupazioneDiCorso1Anno

The reasoner work points out how this new class is a subclass of Occupazione1Anno, which is the class of all the one-year employment survey values; this because, as the question regards only first-level degree courses, having a length of three years, no employment instance of the three or five year employment survey can be found.
The defined class is eventually tested with the addition of an instance having the ISCRITTO_MAGISTRALE property; the inferred class, from the definition of the domain of the property, is OccupazioneMaster, and then every addition to this instance of other property not bound to the Occupazione class will lead to inconsistencies.



Description: OccupazioneMaster

Equivalent To

SubClass Of
Occupazione1Anno

Description: ISCRITTO_MAGISTRALE

Equivalent To

SubProperty Of

Domains (intersection)
OccupazioneMaster

Ranges
xsd:integer

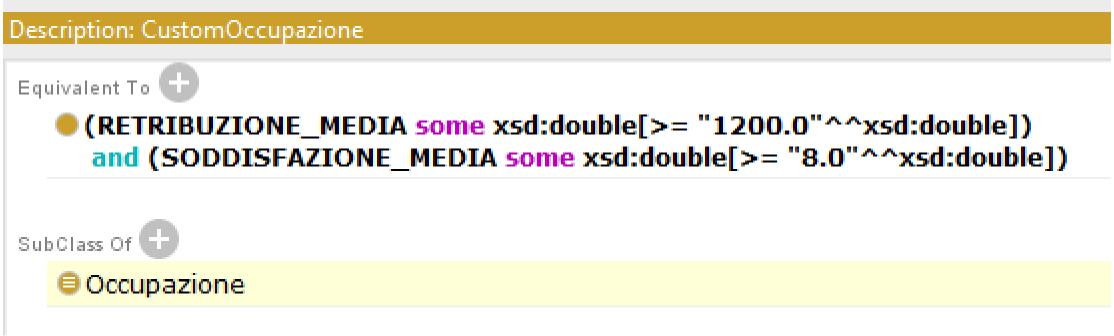
Figure 21 - Newly defined class OccupazioneMaster as domain of ISCRITTO_MAGISTRALE property

Intensional defined classes

A final analysis is made on the classes defined from property restrictions. These classes identify their essence basing on their characteristics, and so the possibility of their definition is useful in order to search particular patterns starting directly from the requirements.
In the defined ontology, an example of custom class of this kind can be the one which represents a Mega University of northern Italy where the released first level degrees give, after one year, an minimum average salary of 1200 euro and have an average satisfaction higher or equal to 7 (out of 10). In order to accomplish this purpose, the work has been separated in two parts: first, the definition of the custom employment class based on the specified characteristics, and after the creation of a university class respecting the custom employment

Custom class for employment

The first class generated has been called *CustomOccupazione*, and is a subclass of Occupazione class which defines specific data properties values. In particular, the data properties involved are RETRIBUZIONE_MEDIA (about the average salary) and SODDISFAZIONE_MEDIA (about the average satisfaction). The expressed property reveals necessary and sufficient conditions for the membership of the class. Figure 22 reveals the class definition in Protégé.



Description: CustomOccupazione

Equivalent To ⊕

```
● (RETRIBUZIONE_MEDIA some xsd:double[>= "1200.0"^^xsd:double])
   and (SODDISFAZIONE_MEDIA some xsd:double[>= "8.0"^^xsd:double])
```

SubClass Of ⊕

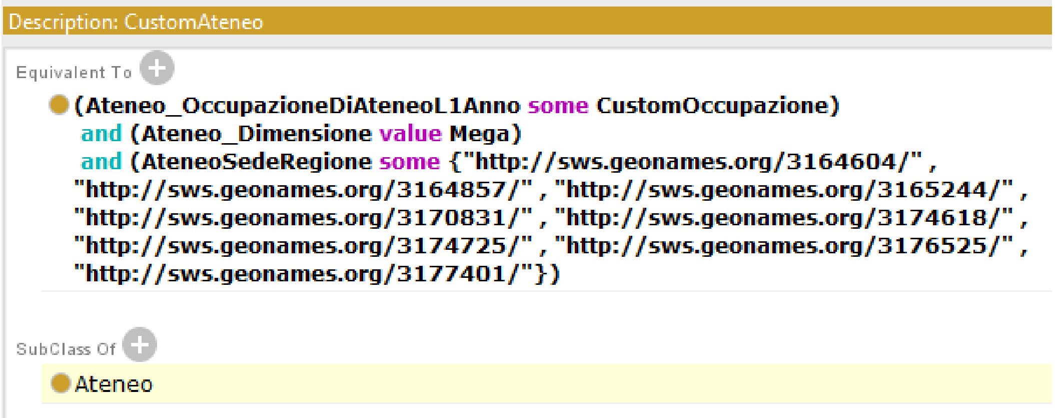
● Occupazione

Figure 22 - Definition of class CustomOccupazione

Custom class for university

Consequent to the definition of the CustomOccupazione class, the class regarding the characteristics of the wanted university has been created. Its name is *CustomAteneo* and involves the properties following properties:

- Ateneo_Dimensione, for the size of the University (Mega is a dimension which means more than 40000 enrolled students)
- AteneoSedeRegione for the definition of the allowed regions (northern Italy in this case)
- Ateneo_OccupazioneDiAteneoL1Anno: object property binding a university to its first-level degree instance of the one-year employment survey. In this case it should be an occurrence of CustomOccupazione type.



Description: CustomAteneo

Equivalent To ⊕

```
● (Ateneo_OccupazioneDiAteneoL1Anno some CustomOccupazione)
   and (Ateneo_Dimensione value Mega)
   and (AteneoSedeRegione some {"http://sws.geonames.org/3164604/" ,
  "http://sws.geonames.org/3164857/" , "http://sws.geonames.org/3165244/" ,
  "http://sws.geonames.org/3170831/" , "http://sws.geonames.org/3174618/" ,
  "http://sws.geonames.org/3174725/" , "http://sws.geonames.org/3176525/" ,
  "http://sws.geonames.org/3177401/"})
```

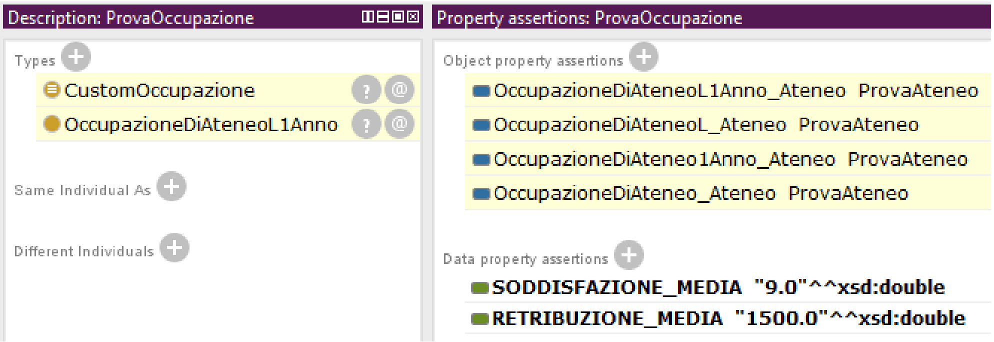
SubClass Of ⊕

● Ateneo

Figure 23 - Definition of class CustomAteneo

Testing of the classes

A way to test the defined class is via the definition of instances in Protégé, so to let the reasoner to make use of them to verify class membership or inconsistencies. The first instance created, called *ProvaOccupazione*, has the data property SODDISFAZIONE_MEDIA set to 9 and the data property RETRIBUZIONE_MEDIA set to 1500. The reasoner in this case infers that this is an instance of the CustomOccupazione class.



Description: ProvaOccupazione

Types

CustomOccupazione

OccupazioneDiAteneoL1Anno

Same Individual As

Different Individuals

Property assertions: ProvaOccupazione

Object property assertions

OccupazioneDiAteneoL1Anno_Ateneo ProvaAteneo

OccupazioneDiAteneoL_Ateneo ProvaAteneo

OccupazioneDiAteneo1Anno_Ateneo ProvaAteneo

OccupazioneDiAteneo_Ateneo ProvaAteneo

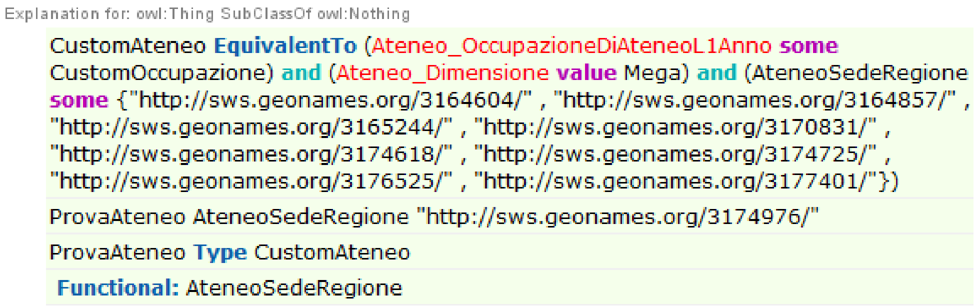
Data property assertions

SODDISFAZIONE_MEDIA "9.0"^^xsd:double

RETRIBUZIONE_MEDIA "1500.0"^^xsd:double

Figure 24 - Instance inferred as CustomOccupazione

Finally it has been created an instance (named *ProvaAteneo*) of a class having Mega as size of the University, linked with the created ProvaOccupazione instance via the Ateneo_OccupazioneDiAteneoL1Anno object property and having a northern Italy region defined via the AteneoSedeRegione data property. The reasoner states that this instance belongs to the class CustomAteneo. In order to make a final verification, one of the properties has been changed: in particular, the region value has been updated to the geonames value of Lazio (3174976). This new configuration does not represent a correct instance for the CustomAteneo class (lack of necessary conditions). Therefore, if with this setting the ProvaAteneo instance is forced to be a member of the CustomAteneo (through the explicit statement), an inconsistency would be reported by the reasoner. Figure 25 reports the final check.



Explanation for: owl:Thing SubClassOf owl:Nothing

```
CustomAteneo EquivalentTo (Ateneo_OccupazioneDiAteneoL1Anno some
CustomOccupazione) and (Ateneo_Dimensione value Mega) and (AteneoSedeRegione
some {"http://sws.geonames.org/3164604/" , "http://sws.geonames.org/3164857/" ,
"http://sws.geonames.org/3165244/" , "http://sws.geonames.org/3170831/" ,
"http://sws.geonames.org/3174618/" , "http://sws.geonames.org/3174725/" ,
"http://sws.geonames.org/3176525/" , "http://sws.geonames.org/3177401/"})
ProvaAteneo AteneoSedeRegione "http://sws.geonames.org/3174976/"
ProvaAteneo Type CustomAteneo
 Functional: AteneoSedeRegione
```

Figure 25 - Inconsistency due to an incorrect value of AteneoSedeRegione data property

Data visualization

The definition of the AlmaLaurea ontology and the following generation of the triple store have been accompanied by the development of a reporting tool which permits the visualization of the data. In addition, a modular interface for the creation of queries and their subsequent launch has been created. In this chapter there are explained all the steps and the motivations which led to the creation of the software.

Graphical reports

Motivations

Since ancient times, humans have found in graphical representations a way to improve the communications of concepts. Among all the treated arguments, the representation of data has brought an important impact in the knowledge understanding, assuring a great efficiency thanks to the exploitation of the human visual perception and cognitive system abilities. The visualization has allowed *explorative analysis* of data, with the purpose to identify their structure, properties and patterns. According to Jacques Bertin, this kind of analysis denotes "the visual instrument to solve logical problems" [77]. Through the visual analysis of the data it is possible to extract information from them, dealing with the *information visualization.* The contribute of the usage of the diagrams then deeply influences the process of *understanding continuum* [78] which steps from data to information, then to knowledge and finally to wisdom (DIKW model).

The visualization of the data is a cognitive process where a person builds a mental model of the data, according to Robert Spence [79]. This implicit work has a great impact on the understanding and the reasoning on the information, and these improvements are object of many researches. Card et al. [80] explain how the graphical representations can help also in the deduction of new information, thanks to the appliance of perceptive inference processes. A study by Larkin and Simon [81] state how the expressiveness of data is more effective with the usage of diagrams, identifying the reasons in three different properties:

- **Localization**: the correct presence of an information plotted in a given space eases the comparison with other data.
- **Minimum labeling**: the similarity of a graphical element with the real world ensures a better understanding with respect to the corresponding textual value.
- **Perceptual enhancement**: many inferences can be effortlessly done when looking at a diagram (e.g. clustering over a given zone)

The explained motivations have guided to the construction of a graphical report tool, described in the following sections.

Basic Idea

Inspired by the existing SUA reports available in the AlmaLaurea's university staff website, the main goal is to create a similar graphical report, providing the same level of knowledge expressivity, starting from the data available in the newly defined open format. Furthermore, to accomplish the openness paradigm, the creation of these tools has been done with the help of open source and freely available software instruments. The full process has also the aim to point out a collateral advantage of the exploitation of the structured open data: not only a powerful tool for automatic semantic reasoning, but also a way to share data among humans in a standard defined style, so that to promote the distribution and the reuse of the information. The visualization of the open data in facts demonstrates an immediate possible way of their reutilization.

Technological stack

The graphical reports development has been done with the purpose of the creation of a web-based software platform, dynamically populated. The relative absence of server-side computation (apart from the RDF triplestore previously explained) has simplified the choice of the instruments for the development, identifying several front-end tools, and guiding the approach to a lightweight work methodology without the use of particular frameworks. Here follows a brief resume of the utilized technological products.

HTML5

The latest version of the language for the creation of web pages has been used as base for the definition of the markup and the backbone of the application.

CSS3

Together with the previous, forms an unyielding couple to style the pages and provide a graceful visualization of the page layout.

Twitter bootstrap

Among the most adopted font-end frameworks, it provides a toolkit of front end features to create web application in a stable and conventional way. From the version 4, the grid layout system is based on the CSS3 Flexbox specification methodology, which has been adopted in the final project GUI implementation to assure the mobile and responsive support.

Javascript

Leading language for the front-end web programming, for the current project covers all the calculus and data representation structures. Thanks to the support of functional constructs, it has permitted to exploit the deep use of recursive approaches, improving the global computation.

jQuery

Javascript open source library which simplifies the syntax for the Document Object Modeling (DOM) navigation and the definition of events handlers. Its utilization is mainly adopted because of the support of the asynchronous Javascript and XML (AJAX) techniques.

Google chart

The Javascript-written graphical library made available by Google has been used as the core of the current software project. The ease of use has permitted a rapid development of the data visualization, obtained with the simple provision of data and the setting of configuration options.

Data retrieving

The data retrieving has been made exploiting a feature of the Fuseki server: thanks to its support of SPARQL Over HTTP (SOH) commands, the queries have been resolved with a set of HTTP requests. In particular, Fuseki exposes a SPARQL endpoint, queryable in a RESTFul way. Once defined the query to be launched, it has been transformed with the application of URL encoding and passed as GET parameters of the URL of the Fuseki endpoint. The possibility of the attainment of the results in Javascript Object Notation (JSON) format ensured the possibility of use from the defined web application. More specifically, with the usage of the jQuery library, each request has been done with AJAX techniques.

First example

In a similar way respect to the SUA existing reports, the development of the new web application for the visualization has been divided into two different sections: one for the profile survey and one for the employment survey. Basing on the characteristics of the questionnaires, different kinds of charts have been adopted. Among the variety of chart types offered by the Google chart library, the following have been chosen:

- **Pie chart:** chosen for the displaying of the variables of the profile survey. Its adoption is mainly due to the fact that it allows to see the whole distribution of the different answer values of each single question of the survey.
- **Column chart:** chosen for the displaying of the variables of the employment survey. This kind of visualization focuses on the numerical comparison of the replies of different questions of the survey.

The development of the first version of the application followed several considerations regarding the knowledge pattern to show, starting from the retrieved RDF data. The main idea in this case is to explain the meaning of the data in a more human-friendly mode. For both the questionnaires, the leading decisions are explained next.

Profile survey

The report about the graduates' profile is divided into ten different questions, each of them having different possible values for the reply. Thus, for each question, a different pie chart has been created. The total on which the profile survey is calculated is the number of graduates interviewed in the last three years. This values equals the sum of all possible values (including the "not responding" option) of each question. Using the pie chart is then possible to see the full distribution of all possible values over a given question.

Employment condition survey

In the case of employment, the six variables of the questionnaire refer to a direct aggregated value. Basing on their type, the available questions have a different referencing total. This further separation has been adopted for the implementation of the reports: questions having the same total are displayed in the same column chart. In the next table are reported the deeper details for each question, specifying which RDF property is related to the question and to the corresponding total:

variable RDF property	Variable	Total	Total RDF property
NUM_OCCUPATI	Gradutes currently employed	Number of graduates not working at degree obtainment	NUM_INTERVISTATI_NONLAVORAVANO
ISCRITTO_MAGISTRALE	Graduates enrolled to a master degree	Number of graduates not working at degree obtainment	NUM_INTERVISTATI_NONLAVORAVANO
NONCERCA_MAFORM	Graduates not working but enrolled to a university or professional course	Number of graduates not working at degree obtainment	NUM_INTERVISTATI_NONLAVORAVANO
UTILIZZO_COMPETENZE	Graduates strongly using competences acquired with their degree	Graduates working after degree obtainment	LAV
RETRIBUZIONE_MEDIA	Average salary	No total (already an average value)	-
SODDISFAZIONE_MEDIA	Average satisfaction for the current employment	No total (already an average value)	-

Dynamic retrieval of course list

The starting point of the visualization report is the degree course. All the queries for the data retrieval are performed starting on this value. In order to create a more dynamical web application, it has been performed an opposite query aiming at retrieving all the possible courses for which a correct survey value exists. This query constructs the graph until a degree course granularity level, returning distinct degree codes values.

Particular conditions to be satisfied concerns the presence or absence of the variables regarding the questions wanted to be shown into the report. In fact, if for a given course the cardinality of the graduates is under threshold, the information about the instance of the survey questions for that course is not present. This is due to the privacy duties explained in previous chapter. The course retrieving SPARQL query is then based on the previously mentioned course graph together with the presence of at least one value for a predicate regarding a question of the questionnaire.

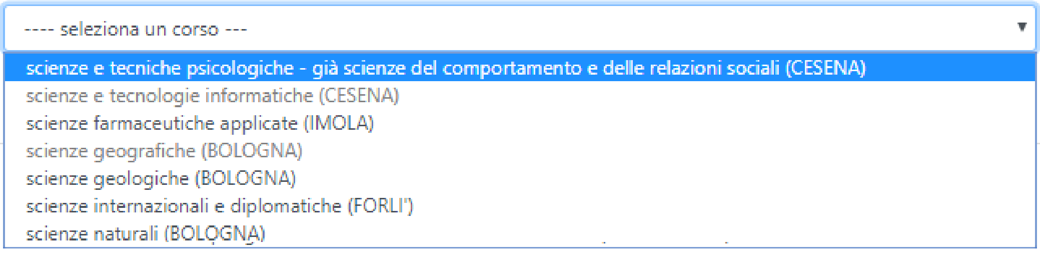
Here follows an example of a query for the retrieval of the list of the first-level degree courses having visible values of the profile questionnaire, basing on the first version of the ontology developed. The presence of path expression on question R105_1 ensures the refinement on the courses not hidden because of privacy matter.

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
PREFIX alma:
<http://www.almalaurea.it/opendata/ontologies/default#>

SELECT DISTINCT ?codicione ?desc ?sede
WHERE {
  ?corso alma:CorsoCodicione ?codicione.
  ?corso alma:CorsoClasseDiLaurea ?subject.
  ?corso alma:CorsoDescrizione ?desc.
  ?subject alma:ClasseDiLaureaTipo ?object.
  ?pro alma:PROFILO_CORSO ?corso.
  FILTER(?object="LT").
  ?pro alma:R105_1 ?R105_1
}
```

While there are courses which satisfy the minimum threshold requirements for data visualization, other courses which do not have the values for the questions are present. To retrieve them, it suffices to change the last path expression, in order to look for all the courses which do not respect the presence of replies for questions of the survey. The graph pattern which replaces the last one is the following:

```
FILTER NOT EXISTS{
  ?pro alma:R105_1 ?R105_1
}
```

The merge of the results of both the previous queries creates a full list of courses, either available or not for visualization. This has been used for the population of a dropdown for the selection of the course data to visualize. In figure 26 it is reported the result of this process in the web application, containing available courses and disabled ones.



---- seleziona un corso ---

scienze e tecniche psicologiche - già scienze del comportamento e delle relazioni sociali (CESENA)
scienze e tecnologie informatiche (CESENA)
scienze farmaceutiche applicate (IMOLA)
scienze geografiche (BOLOGNA)
scienze geologiche (BOLOGNA)
scienze internazionali e diplomatiche (FORLI')
scienze naturali (BOLOGNA)

Figure 26 - Example of implementation of the course dropdown within the application

Update after ontology clarification

Together with the first clarification of the ontology, the visualization tool has been modified and adapted to accomplish the newly defined classes and properties of the updated ontologies. While the idea of the reports has remained unaltered, with the representation of the same questions in the charts according to the previous considerations, the retrieving queries have been modified. The performed modifications have also allowed the visualization of all the three different employment surveys results, reporting the values of each single question in paired columns in the chart. Figure 27 shows the visual outcome for a question.

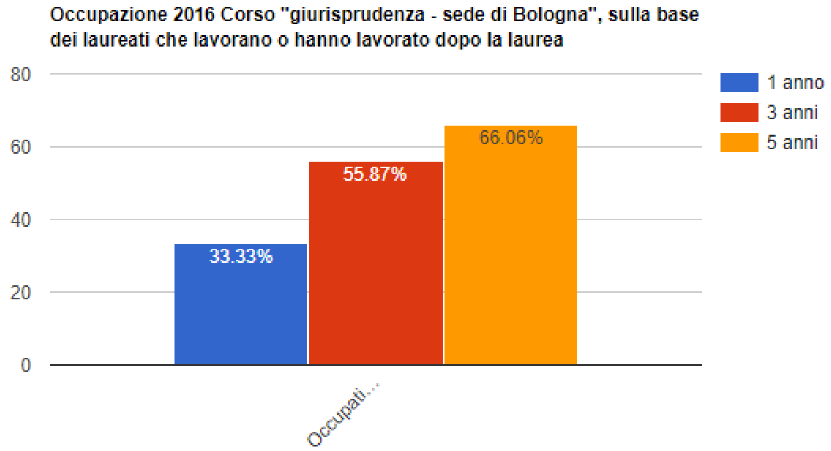


Figure 27 - Example of multi-year employment survey question visualization

The described update has also influenced the construction of the course list. Particular changes have been done for the employment survey, because of the fact that a course can have valid values for not all the three existing surveys. The list of available courses has been then constructed by including all the possible ones who have at least one valid survey datum: the graph pattern has not been bound to a specific object, but to all the possible object which satisfy a triple having a particular predicate (referring to one, three or five year survey). This has been done exploiting the hierarchy property stated into the ontology via the *rdfs:subPropertyOf* predicate. As an example, a new property has been defined in the SPARQL query, as sub property of alma:Corso_OccupazioneDiCorso. By looking for triples which satisfy this predicate, the query retrieves all the courses satisfying the properties alma:Corso_OccupazioneDiCorso1Anno, alma:Corso_OccupazioneDiCorso3Anni and alma:Corso_OccupazioneDiCorso5Anni.

Moreover it has been modified the mechanism of retrieval of the non-available courses: a course is considered not available if it has not values for none of the three surveys. The updated graph pattern, conforming to the new ontology, is the following:

```
?corso alma:CorsoOccupazione1anno2016 ?occ
FILTER NOT EXISTS{
  ?occ occupazione:NUM_INTERVISTATI_NONLAVORAVANO
?nonlavoravano
}.
?corso alma:CorsoOccupazione3anni2016 ?occ3
FILTER NOT EXISTS{
  ?occ3 occupazione:NUM_INTERVISTATI_NONLAVORAVANO
?nonlavoravano
}.
?corso alma:CorsoOccupazione5anni2016 ?occ5
FILTER NOT EXISTS{
  ?occ5 occupazione:NUM_INTERVISTATI_NONLAVORAVANO
?nonlavoravano
}
```

Cohort comparison

As completion of the development of the software, it has been added the possibility to compare the data of a given course with the information coming from the related university or degree class collective. In a similar way to the existing SUA reports (as stated in figure 13), each chart of a survey variable for a given course is placed side by side to the correspondent one of the chosen collective, allowing the comparison of the results. The JavaScript construction of the charts is similar, as it differs only on the same instantiation of different objects fed with different data. Besides, the queries defined for the extraction of the data are similar nevertheless the cohort to extract (either course, university or degree class). An example of SPARQL query used to retrieve the 1-year employment survey data of a first level degree class is the following:

```
PREFIX occ:
<http://www.almalaurea.it/opendata/ontologies/occupazione#>
PREFIX alma:
<http://www.almalaurea.it/opendata/ontologies/default#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT
(xsd:integer(?NUM_OCCUPATI) /
xsd:integer(?NUM_INTERVISTATI_NONLAVORAVANO) as
?NUM_OCCUPATI_AVG)
((xsd:integer(?ISCRITTO_MAGISTRALE)) /
(xsd:integer(?NUM_INTERVISTATI_NONLAVORAVANO)) as
?ISCRITTO_MAGISTRALE_AVG)
((xsd:integer(?NONCERCA_MAFORM)) /
(xsd:integer(?NUM_INTERVISTATI_NONLAVORAVANO)) as
?NONCERCA_MAFORM_AVG)
((xsd:integer(?UTILIZZO_COMPETENZE))/ (xsd:integer(?LAV)) as
?UTILIZZO_COMPETENZE_AVG)
(?RETRIBUZIONE_MEDIA as ?RETRIBUZIONE_MEDIA_AVG)
(?SODDISFAZIONE_MEDIA as ?SODDISFAZIONE_MEDIA_AVG)
WHERE {
  ?occ occ:OccupazioneDiClasseL1Anno_Classe ?classe.
  ?classe alma:ClasseDiLaureaCodice ?classeCodice.
  FILTER(?classeCodice='10040')
  ?occ occ:ANNO_INDAGINE ?ANNO_INDAGINE.
  ?occ occ:NUMLAU ?NUMLAU.
  ?occ occ:NUMLAU_NONLAVORAVANO ?NUMLAU_NONLAVORAVANO.
  ?occ occ:NUM_INTERVISTATI ?NUM_INTERVISTATI.
  ?occ occ:NUM_INTERVISTATI_NONLAVORAVANO
?NUM_INTERVISTATI_NONLAVORAVANO.
  ?occ occ:NUM_OCCUPATI ?NUM_OCCUPATI.
  occ:NUM_OCCUPATI rdfs:comment ?NUM_OCCUPATI_DESC.
  ?occ occ:ISCRITTO_MAGISTRALE ?ISCRITTO_MAGISTRALE.
  occ:ISCRITTO_MAGISTRALE rdfs:comment
?ISCRITTO_MAGISTRALE_DESC.
  ?occ occ:NONCERCA_MAFORM ?NONCERCA_MAFORM.
  occ:NONCERCA_MAFORM rdfs:comment ?NONCERCA_MAFORM_DESC.
  ?occ occ:UTILIZZO_COMPETENZE ?UTILIZZO_COMPETENZE.
  occ:UTILIZZO_COMPETENZE rdfs:comment
?UTILIZZO_COMPETENZE_DESC.
  ?occ occ:LAV ?LAV.
  ?occ occ:RETRIBUZIONE_MEDIA ?RETRIBUZIONE_MEDIA.
  ?occ occ:SODDISFAZIONE_MEDIA ?SODDISFAZIONE_MEDIA.
}
```

The creation of the web reports has been done contemporary to the final modification of the ontologies made to let the collective comparison; the different classes defined in the ontologies have caused a fragmentation of the development of the reports, leading to four different versions of the software for each survey type:

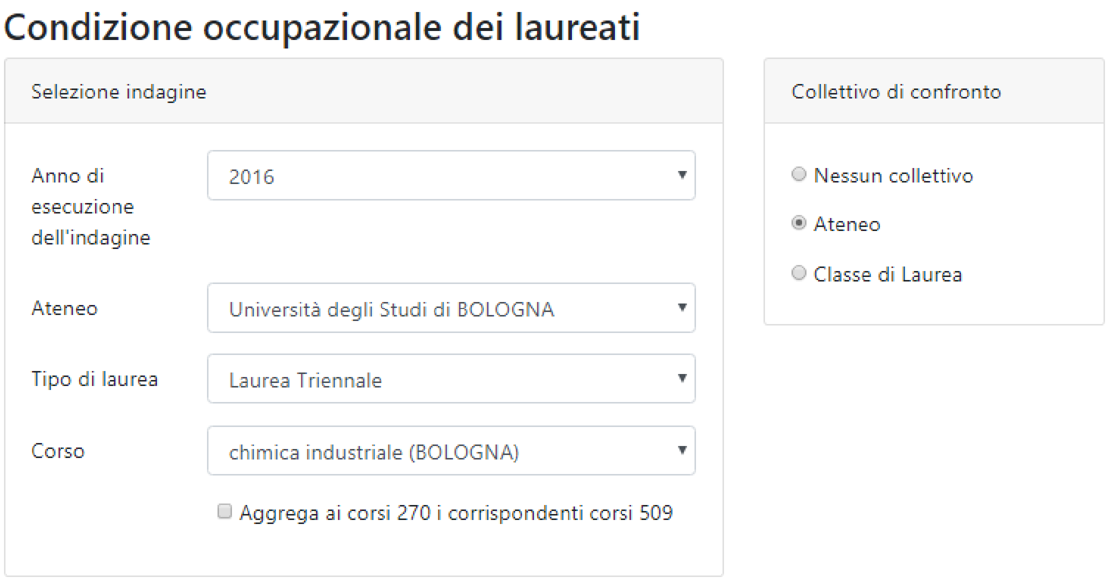
- Single degree data compared to university data
- Aggregated single degree data (considering also the previous version of the courses ruled by D.M. 509/99) compared to university data
- Single degree data compared to degree class data
- Aggregated single degree data compared to aggregated degree class data (both considering also the previous version of the courses ruled by D.M. 509/99)

Final unified interfaces

Pursuing the simplification of the instrument and its usage improvement, the final software created consists in two reports, one for each survey type. These reports derive from the four previously defined, whose behaviors have been unified in a single interface, allowing a global data visualization experience through rapid switches on the collectives to discover. The proposed report contains a parametric form that consents the choice of the degree course on which to perform the data visualization. Notably, it has been defined a set of conditional dropdowns which filter the full list of courses. The dropdowns regard the following criteria:

- Year of execution of the survey
- University
- Type of degree (level)

Similarly to the construction of the dropdown of the courses, also these ones have been populated starting from the RDF triplestore, by executing a query via the HTTP restful endpoint. The form is then completed with the presence of other input controls which allow the change of the cohort for the comparison and the inclusion of aggregated values. Figure 28 reports the final aspect of the visualization form.



Condizione occupazionale dei laureati

Selezione indagine

Anno di esecuzione dell'indagine	2016
Ateneo	Università degli Studi di BOLOGNA
Tipo di laurea	Laurea Triennale
Corso	chimica industriale (BOLOGNA)

□ Aggrega ai corsi 270 i corrispondenti corsi 509

Collettivo di confronto

○ Nessun collettivo

● Ateneo

○ Classe di Laurea

Figure 28 - Parametric form for the visualization of data about employment condition's survey

Eventually, to maintain a better truthfulness of the information shown, several warnings about possible data inconsistencies have been implemented. The particular cases managed regard:

- **Inter-class courses**: when showing a comparison over the degree class for these types of degrees a warning is shown, because the degree class data refers to the aggregation of all the possible single degree class values.
- **Previous / next version of a course**: implemented as fallback in the aggregated data mode of the employment report, this information is used to retrieve data of previous version of a course when there is an empty result of the current course.

A wizard for query building

In order to guarantee the full utilization of the open knowledge base generated, another software implemented into the thesis project regards a wizard interface for the incremental creation of SPARQL queries, to be launched on the available Fuseki endpoint. The idea of the tool has been inspired by other more known examples, like the ISTAT open data query construction platform or the European Data Portal Linked data query wizard.

The reasons of the work consist in letting the non-technical users able to perform specific data retrieving requests, exploiting the structure defined by the created ontologies.

Interface design

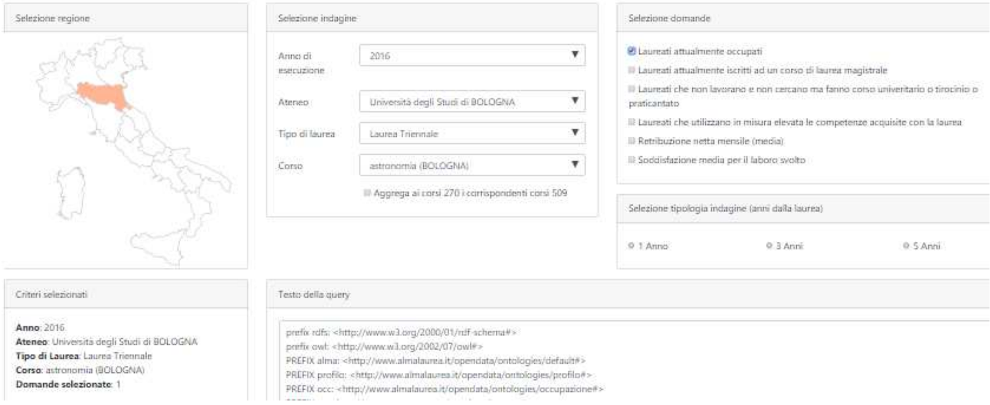
The software, divided in two interfaces basing on the different survey to be queried, aims at the growing granular construction of SPARQL queries, starting from the university level until the degree course level, focusing on the institution didactic hierarchy instead of the degree class one. The proposed interface, built with the same technologies used for the data visualization platform, is combined by three sections, each containing different aspects for the query construction:

- **Geographical map**: used to refine the dataset basing on the establishment region of the university.
- **Survey / degree choice**: same set of filters present in the data visualization software, consists on performing year of survey, degree choice and aggregated visualization option.
- **Variables to be extracted**: a series of checkboxes is listed, corresponding to the existing variables of the survey. The choice of a variable includes its possible values into the SPARQL query.

In the employment condition survey, also a fourth box is showed, regarding **the choice of the kind of employment survey by year**, for filtering the results in one of the three available employment surveys.

Basing on the configuration of the form, several scenarios of non consistency can happen. Therefore, the implemented software considers also the validity of the chosen combination of parameters, pledging the creation of valid queries. As an example, within the employment survey, the form disables the selection of the question “graduates currently enrolled to a master degree course” if the selected degree type is not a bachelor level. This because that question is present only for the first-level degrees surveys, and the inclusion of the graph pattern in other degree types would lead to an empty result.

The described filter boxes are followed by a parameters recap box and by a textarea where the generated query is present, ready to be launched. On figure 29 is reported the resulting form.



Selezione regione

Selezione indagine

Anno di esecuzione: 2016

Ateneo: Università degli Studi di BOLOGNA

Tipo di laurea: Laurea Triennale

Corso: astronomia (BOLOGNA)

□ Aggrega ai corsi 270 i corrispondenti corsi 509

Selezione domande

☑ Laureati attualmente occupati
□ Laureati attualmente iscritti ad un corso di laurea magistrale
□ Laureati che non lavorano e non cercano ma fanno corso univeritario o tirocinio o praticantato
□ Laureati che utilizzano in misura elevata le competenze acquisite con la laurea
□ Retribuzione netta mensile (media)
□ Soddisfazione media per il laboro svolto

Selezione tipologia indagine (anni dalla laurea)

○ 1 Anno ○ 3 Anni ○ 5 Anni

Criteri selezionati

Anno: 2016
Ateneo: Università degli Studi di BOLOGNA
Tipo di Laurea: Laurea Triennale
Corso: astronomia (BOLOGNA)
Domande selezionate: 1

Testo della query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
PREFIX alma: <http://www.almalaurea.it/opendata/ontologies/default#>
PREFIX profilo: <http://www.almalaurea.it/opendata/ontologies/profilo#>
PREFIX occ: <http://www.almalaurea.it/opendata/ontologies/occupazione#>
```

Figure 29 - Employment condition query wizard form

Query construction logic

The designed form allows the interrogation over the dataset at different granularity levels. According to the ontology structure, the different levels correspond to different objects instances of subclasses of the main classes representing the survey values. Thus, to maintain the correctness of the result retrieval, different possible paths have been individuated, produced by the possible combinations of filters in the form. The possibility to act on the hierarchy of concepts is given by the definitions present into the ontologies. In particular, the queries have been constructed including the checking of properties searched using the *rdf:subPropertyOf* property.
Apart from the variable choice, which is independent from the collective selection, the different cohort selection is based on the presence of different graph patterns in the WHERE clause of the SPARQL query. Here follows a brief analysis of the possible main filtering selections for the query building.

Choice of survey year and region

The wider level of granularity predicate, applicable for both the surveys, returns all the values of the University of a region for all the possible degree types (e.g. for the profile survey, all the instances which satisfy the sub properties of *ProfiloDiAteneo_Ateneo*, which connects a university profile survey value with its university object). The filtered predicates are the year of each survey value (data property named ANNO_INDAGINE) and the geonames region code of establishment of the university.

Choice of survey year and university

A more specific restriction based on both the survey year and the university.

Even in this case the retrieval is based on the sub property as described in the previous option, while the filtered predicates are about the year of the survey value and the university code. As this filter is more detailed respect to the region one, if chosen together it replaces the region filter,

Choice of survey year, university and type of degree

This case refines beyond the previous one. The choice of the degree type specifies the kind of predicate to be analyzed, so in this scenario it is not needed to look for sub properties of a given one, but it is possible to directly search for the specific predicate. Updating the previous example, in case of selection of first level degree type, the instances to be found must be connected to the property *ProfiloDiAteneoL_Ateneo* (which is the sub property of *ProfiloDiAteneo_Ateneo* for the first level degrees only). Moreover, the FILTER clauses on survey year and university code remain the same.

Choice of survey year and degree course (profile survey)

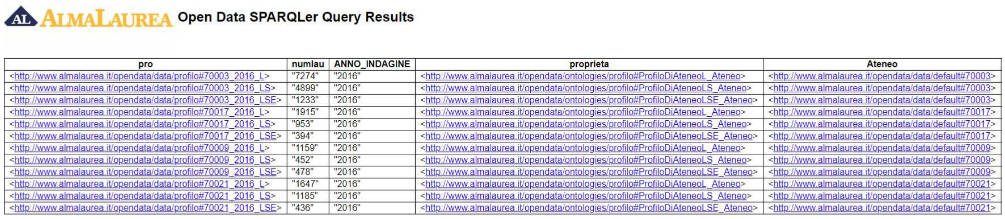
The most granular level of search, is based not on the search of university profile survey value, but on more specific single degree profile survey value. The searched property in this case is *ProfiloDiCorso_Corso*, which connects the searched profile instance to the related course instance. The applied FILTER clauses within the query are about the survey year and the degree code value.

Choice of survey year and degree course (employment survey)

Like in the same filtering situation of the profile survey, the settings on objects and predicates to find are identical. In the case of employment survey it is however necessary to consider the presence of three different surveys, fact already mentioned and implemented into the ontology and the RDF generation. For this reason, even in this case the searched property must be supported by the *rdfs:subProperty*. For instance, the search of values of a given course is done by looking for sub properties of *OccupazioneDiCorso_Corso:* in this way all the values for the three different survey types (one, three and five years) are returned. The considerations on the presence of the survey year dimension within the employment condition survey form have been applied also to all the other scenarios.

Results

After the parameterization of the form, the final step is the launch on the query on the Fuseki server. This has been done exploiting the same HTTP endpoint used for the AJAX calls performed in the data visualization software. In this case, instead of obtaining the results in JSON, the chosen format is XML. Thank to the application of layout rules defined in extensible stylesheet language (XSL), the result appear in a styled tabular design. The result of a query launch is visible in figure 30.



Open Data SPARQLer Query Results

pro	numlau	ANNO_INDAGINE	proprieta	Ateneo
<http://www.almalaurea.it/opendata/data/profilo#70003_2016_L>	"7274"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoL_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70003>
<http://www.almalaurea.it/opendata/data/profilo#70003_2016_LS>	"4899"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLS_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70003>
<http://www.almalaurea.it/opendata/data/profilo#70003_2016_LSE>	"1233"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLSE_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70003>
<http://www.almalaurea.it/opendata/data/profilo#70017_2016_L>	"1915"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoL_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70017>
<http://www.almalaurea.it/opendata/data/profilo#70017_2016_LS>	"953"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLS_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70017>
<http://www.almalaurea.it/opendata/data/profilo#70017_2016_LSE>	"394"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLSE_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70017>
<http://www.almalaurea.it/opendata/data/profilo#70009_2016_L>	"1159"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoL_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70009>
<http://www.almalaurea.it/opendata/data/profilo#70009_2016_LS>	"452"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLS_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70009>
<http://www.almalaurea.it/opendata/data/profilo#70009_2016_LSE>	"478"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLSE_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70009>
<http://www.almalaurea.it/opendata/data/profilo#70021_2016_L>	"1647"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoL_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70021>
<http://www.almalaurea.it/opendata/data/profilo#70021_2016_LS>	"1185"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLS_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70021>
<http://www.almalaurea.it/opendata/data/profilo#70021_2016_LSE>	"436"	"2016"	<http://www.almalaurea.it/opendata/ontologies/profilo#ProfiloDiAteneoLSE_Ateneo>	<http://www.almalaurea.it/opendata/data/default#70021>

Figure 30 - Results of the query launch

A final note regards the efficiency of performances of the engine: due to the enormous size of the dataset, loose queries can lead to very long execution time, until a stuck situation in the browser rendering. For this reason, the minimum detail level allowed for the guided queries is the combination of both survey year and university region. A message into the form warns about the application of this policy.

Final considerations

Environment Evaluation

An evaluation of the work has been done basing on four different quality criteria: usability, portability, availability and performance. For all the different aspects the evaluation is done on the RDF dataset not published yet.

Usability

Regarding the usability, a distinction is done between the usability of publishing organization (AlmaLaurea) and the usability of the consumers of the data.

Usability for publishing organization

Different aspects have to be analyzed for the evaluation of the usability of the publishing organization. First of all, the needed know-how for the deployment of the platform. The technical support for the current project is made of several tools, including Apache Jena Fuseki, Protégé and the TSQL language, used for the definition of stored procedure necessary for the creation of the RDF triplestore. Moreover, for the modification of the data visualization tools, knowledge on the main front-end instruments adopted is needed. Because of the general purpose nature of the tools used, it is trivial for an IT staff member to manage the developed products, and so the first usability requirement is accomplished.

A second aspect regards the usability for the human resources: in this case, as the know-how about the construction of the ontology, of the reporting tools and of the surveys are kept by different people inside the organization, a work of formation is needed: the promotion of seminars and specific trainings about the project can fill the knowledge gap, also because no specific technical knowledge is needed for the use of the ended product.

A final aspect of usability concerns the learning curve: due to the fact that the employees should attend courses for the learning of the platform, the learning curve can result a bit steep. This also because many employees can be non familiar with the concept of RDF, SPARQL and Linked Data.

In conclusion, the producer usability of the AlmaLaurea open dataset is evaluated as satisfied over the three different aspects analyzed.

Usability for data consumer

In compliance with the open paradigm, the structured data proposed are freely available and unconstrained in proprietary applications. The usability of the data for the consumers is guaranteed by the definition of the data visualization tools, which allows the user to retrieve graphical information about the knowledge in the dataset. The additional wizard interface for the SPARQL query creation and the consequent possibility of query and result download assure a full availability.

Portability

To evaluate the portability there have been analyzed again different aspects. First, the environment openness: Despite the most part of the utilized tools are open source, the principal source of data to be reified in RDF/XML format is stored on a Microsoft SQL server database, with a proprietary license. Focusing on another aspect, which are the possible external dependencies with the AlmaLaurea environment, the produced dataset results independent, even if the definition of the ontologies follow completely the indications of the surveys published by the consortium, and so possible structure modifications can happen in future releases. For these reasons, the portability of the environment results limited, and a new version of the triplestore generation software should be redefined from scratch.

Availability

The availability of the data preparation environment is actually stuck at the current existing dataset. In future, as the AlmaLaurea consortium releases annually the data about the performed surveys, an extraction to generate the RDF format can be executed every year, possibly adapting or updating the model of data according to possible modifications happened. Moreover, the availability of a public SPARQL server environment has to be guaranteed.

Performance

The evaluation of performance can be analyzed on two different aspects: the time needed for the generation of the triplestore and the throughput of the data visualization tools. The first aspect can be ignored, as it strongly depends on the computational power of the SQL server database machine. Moreover, as the extraction is not frequent (once a year) it is not a problem if the execution takes several hours. Regarding the data visualization tools, most of the responsibilities for the performances are prerogative of Apache Jena Fuseki server and JavaScript optimized code, while the first difficulty scales with the growth of the number of the triples over a single scheme, for the second case an important improvement have been noticed thanks to the usage of functional constructs of JavaScript. A satisfying benchmark is anyway guaranteed with the generation of all the survey data for a 3-year period.

Final product counts

The finished project has led to the definition of many concept. In particular, the results of the ontology creation consist of:

- 74 classes
- 121 object properties
- 107 data properties
- 1628 total axioms

Regarding the reified RDF triplestore dataset, it is formed by 3161153 distinct triples.

Conclusions

The presented thesis project aims at the definition of a referencing ontological model for the description of the statistics on Italian graduates. Through the described steps different possible real cases have been analyzed, and the resulting ontology constitutes a careful starting point for the development of the formal definition of the domain. Thanks to the full usage of the OWL features, the decided modeling has been also confirmed by the feedback of an automatic reasoner.

Given the current growth of the open data movement within the public administration field, the created ontologies make use of the AlmaLaurea survey’s data to define a conceptualization of the graduate’s statistics field that aspires to become a quasi-standard for the description of the domain; this desire is supported by the relative absence of similar models at an international level, whereas instead constitute a leading example within the Italian university panorama.

The parallel development of two different graphical interfaces in addition to the definition of the models has the dual purpose of helping people to exploit better the newly created structured data. Indeed, this consideration, apart from few similar examples, comes from the fact that much of the open data available on the internet is not supported by visualization tools, aspect that often causes a poor consideration of them and leads to a discontinuance of their maintenance. A visualization tool and a query wizard tool help to bridge the gap of the usage of the structured data also for non technical users, making the work usable from a 360-degree point of view.

Beyond the definition of the model, the generation of the structured data represents an important contribute to the world of open data: due to the uniqueness of the kind of information treated, its usage by third part organization can result significant in order to improve a global knowledge about the graduates and university domains. The natural continuation of the project, which consists in the integration of all the other variables present in the questionnaires, can further improve the scenario.

Possible future scenarios of the usage of the data can be hypothesized according to the existing open dataset of the educational domain, and with others regarding the targeted job placement of the graduates, like the open data released by the Italian Ministry of Labour and Social Policy. This opportunity goes towards the direction indicated by the same ministry, which through its job portal *ClicLavoro* has promoted the open data as “engine of the European Union’s economy” [82]. Moreover, the integration with international open datasets like the ones exposed by the European Union open data portal can help to compare the Italian graduates’ performances with those from others countries, process which leads to an increase of the knowledge in the domain by providing a simple benchmark.

Apart from the integration of external datasets, other possible patterns can be investigated by exploiting the existing structured information. A first example can be the extraction of a time series that reports the different performances of the graduates in given courses over the years. Another example can be a comparison of the universities results basing on their dimension (e.g. comparison of the engineering graduates’ performances between a small and a mega university). A final example can regard the comparison between northern and southern university, basing on the region values stored in the data. This last one, still actual, can help the institutions and the universities to verify the causes of the differences in the performances, with possible significant reflexives on the data knowledge and on the economical and social growth of the country.

References

[1] T. Berners-Lee, *Information Management: A Proposal* , W3C, March 1989, https://www.w3.org/History/1989/proposal.html

[2] D.Connolly, A *Little History of the World Wide Web*, W3C, 1995, https://www.w3.org/History.html

[3] T. O'Reilly, *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*, Munich Personal RePEc Archive, March 2007

[4] T.Berners-Lee, J. Hendler and O.Lassila, *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, Scientific American, May 2000

[5] T.Berners-Lee, *Semantic Web - XML2000*, slide 10, W3C , 2000

[6] L. Masinter, T. Berners-Lee, and R. T. Fielding. *Uniform resource identifier(uri): Generic syntax.*, 2005.

[7] U. Consortium et al., *The Unicode Standard, Version 2.0*, Addison-Wesley Longman Publishing Co., Inc., 1997.

[8] T.Bray, J.Paoli et al., *XML 1.0 Specification*, W3C recommendation, 26 November 2008, https://www.w3.org/TR/REC-xml/

[9] L.R.E. Quin, *Schema*, W3C, 2015, https://www.w3.org/standards/xml/schema

[10] F. Manola, E. Miller, B. McBride, et al. *Rdf primer.* W3C recommendation, 10 February 2004, https://www.w3.org/TR/rdf-primer/

[11] D.Brickley, R.V. Guha*, RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Working Draft, 30 April 2002, https://www.w3.org/2001/sw/RDFCore/Schema/200203/

[12] D. L. McGuinness, F. Van Harmelen, et al., *Owl web ontology language overview.*, W3C recommendation, 10 February 2004, https://www.w3.org/TR/owl-features/

[13] D. Beckett and B. McBride, *Rdf/xml syntax specification (revised)*, W3C recommendation, 10 February 2004, https://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/

[14] T. Berners-Lee, Connolly, D. et al., *N3Logic: A logical framework for the World Wide Web*, Cambridge University Press, 2008

[15] D. Beckett, T. Berners-Lee et.al,. *Turtle-terse rdf triple language*, W3C Team Submission, 2008.

[16] G. Carothers and A. Seaborne., *Rdf 1.1 n-triples*, W3C recommendation, 25 February 2014, https://www.w3.org/TR/n-triples/.

[17] M.Sporny, D.Longley et al., *JSON-LD 1.0. A JSON-based Serialization for Linked Data*, W3C recommendation, January, 2014.

[18] E. Prud, A. Seaborne., *Sparql query language for rdf*, W3C recommendation, January 2008.

[19] Microformat community, *Microformats Wiki*, April 2017, http://microformats.org/wiki/Main_Page

[20] D. Connolly, *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*, W3C Recommendation, 11 September 2007, https://www.w3.org/TR/grddl/

[21] M.Sporny, *RDFa Lite 1.1 - Second Edition*, W3C recommendation, 17 March 2015, http://www.w3.org/TR/rdfa-lite/

[22]WHATWG Consortium, *HTML Living Standard – Microdata*, chapter 5, 9 February 2018, https://html.spec.whatwg.org/multipage/microdata.html

[23] Kavi Goel, Ramanathan V. Guha, and Othar Hansson: *Introducing Rich Snippets*, Google Webmaster Central Blog, 12 May 2009 , https://webmasters.googleblog.com/2009/05/introducing-rich-snippets.html

[24] Jennifer Kyrnin, *Why Use Semantic HTML?*, ThoughtCo website, 5 July 2017, https://www.thoughtco.com/why-use-semantic-html-3468271

[25] K.J. Laskey, K.B. Laskey et al., U*ncertainty Reasoning for the World Wide Web*, W3C incubator group final report, march 2008

[26] T.Berners-Lee : *Linked Data - Design issues*, W3C, 2006

[27] T.Berners-Lee, *Linked Data*, slide 10, "The Great Unveiling" TED presentation, 2009, www.w3.org/2009/Talks/0204-ted-tbl

[28] R.Lehmann, R.Isele et al., *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*, 2013

[29] Open Knowledge Foundation, *Linked Open Vocabularies ,* 2018, http://lov.okfn.org/dataset/lov/

[30] Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak, *Linking Open Data cloud diagram 2017*, 2017, http://lod-cloud.net/

[31] R.Cyganiak, C.Bizer, *Pubby A Linked Data Frontend for SPARQL Endpoints*, Freie Universität Berlin, 2007, http://wifo5-03.informatik.uni-mannheim.de/pubby/

[32] European Parliament, Directive 2003/98

[33] Agenzia per l'Italia Digitale, *Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico,* 2016

[34] Istituto nazionale di statistica, *Linked Open Data*, 2015, http://datiopen.istat.it/documentazione.php

[35] Istituto Superiore per la protezione e la ricerca ambientale, *I Linked Open Data dell'Istituto Superiore per la Protezione e la Ricerca Ambientale*, 2015 http://dati.isprambiente.it/

[36] T.R. Gruber, *A translation approach to portable ontology specifications*, Elsevier, June 1993

[37] W. Pidcock, *What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model?*, Metamodel website, 2003

[38] A.Gilchrist, *Thesauri, taxonomies and ontologies – an etymological note*, CURA consortium and TFPL Ltd, 2002

[39] R. Van Rees, *Clarity in the usage of the terms ontology, taxonomy and classification*, Proceedings of the 2003 cib w78 conference, July 2008

[40] W. Wong, W.Liu, M. Bennamoun, *Ontology Learning from Text: A Look Back and into the Future*, ACM Computing Surveys – CSUR, 2011

[41] A. J. Warner, *A Taxonomy Primer*, Lexicon website, 2002, https://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html

[42] A. Mathes, *Folksonomies – Cooperative Classification and Communication Through Shared Metadata [Online Report]*, Journal of Computer-mediated Communication - JCMC, 2004

[43] H. Rutten, S. Rogers, *Lore: language-based expertise mining*, in B. Stanford-Smith and E. Chiozza, editors, *E-work and E-commerce*, 2001

[44] D. L. McGuinness, *Ontologies come of age*, in D. Fensel, J. Hendler, H. Lieberman and W. Wahlster,editors, *Spinning the semantic web: bringing the world wide web to its full potential.*, MIT press, 2002.

[45] International Organization for Standardization, *ISO 25964-1:2011*, 2-62, ISO Online Browsing Platform, 2011

[46] N. Guarino, *Formal Ontology and Information Systems*, Proceedings of the 1st International Conference, IOS Press, June 1998

[47] T. Slimani, *A Study on Ontologies and their Classification*, In *Recent Advances in Electrical Engineering and Educational Technologies*, INASE, 2014

[48] A. Gómez-Pérez, O. Corcho, *Ontology Languages for the Semantic Web*, In *IEEE Intelligent Systems*, vol. 17, pag. 58, 2002

[49] M. Gruninger et al., *Ontology Summit 2007 – Ontology, taxonomy, folksonomy: Understanding the distinctions*, Appl. Ontology, IOS Press, 2008

[50] Protégé community, *Getting Started with Protege Desktop Editor*, chapter Reasoning, 2018, https://protegewiki.stanford.edu/wiki/Protege4GettingStarted#Reasoning

[51] Xiao Hang Wang et al., *Ontology Based Context Modeling and Reasoning using OWL*, Pervasive Computing and Communications Workshops, Proceedings of the Second IEEE Annual Conference on. IEEE, 2004

[52] T. Kauppinen, J. Trame, A. Westermann, *Teaching Core Vocabulary Specification*, LinkedScience. org, Tech. Rep., 2012

[53] S. Mitchell et al., *The VIVO ontology: enabling networking of scientists*, 2011

[54] VIVO website, *About VIVO*, 2018, http://vivoweb.org/info/about-vivo

[55] R. Styles , N. Shabir, *Academic Institution Internal Structure Ontology*, Talis Information Ltd., 2008, http://vocab.org/aiiso/schema

[56] A. Ameen et al., *Creation of Ontology in Education Domain*, in *Technology for Education (T4E),* 2012 IEEE Fourth International Conference on Technology for Education, 2012

[57] D. Dicheva et al., *Ontological web portal for educational ontologies*, SW-EL'05: Applications of Semantic Web Technologies for E-Learning, 2005

[58] Bologna Process' website, *How does the Bologna Process work?*, 2018, http://archive-2010-2015.ehea.info/article-details.aspx?ArticleId=5

[59] G. Demartini et al., *The Bowlogna ontology: Fostering open curricula and agile knowledge bases for Europe's higher education landscape.*, Semantic Web 4.1, 2013

[60] AlmaLaurea website, *Indagini e ricerche*, 2018, http://www.almalaurea.it/universita/statistiche

[61] ISTAT website, *I laureati e il lavoro*, Note metodologiche, 8 June 2012

[62: ISTAT, *Annuario statistico italiano 2017*, Chapter 7, pag.220, 2017

[63] ISTAT website, *Microdati ad uso pubblico*, 2018, https://www.istat.it/it/archivio/microdati+ad+uso+pubblico

[64] Italian Ministry of Education Statistical office, 2018, http://ustat.miur.it/dati/

[65] EUROSTAT statistics explained website, *Employment rates of recent graduates*, 2018, http://ec.europa.eu/eurostat/statistics-explained/index.php/Employment_rates_of_recent_graduates

[66] EUROSTAT statistics explained website, *Tertiary education statistics*, 2018, http://ec.europa.eu/eurostat/statistics-explained/index.php/Tertiary_education_statistics

[67] EUROSTAT statistics explained website, *Increasing attainment at tertiary level*, in *Tertiary education statistics*, 2018, http://ec.europa.eu/eurostat/statistics-explained/index.php/Europe_2020_indicators_-_education#Increasing_attainment_at_tertiary_level

[68] European Data Portal website, *Education: Open Data in Schools*, 2018, https://www.europeandataportal.eu/highlights/open-data-schools

[69] G. Pirrotta, *Linking Italian university statistics.*, Proceedings of the 6th International Conference on Semantic Systems, ACM, 2010

[70] AlmaLaurea website, *Almalaurea 2016 graduates' profile methodological notes*, 2018, http://www2.almalaurea.it/cgi-php/universita/statistiche/note-metodologiche.php?lang=en&config=profilo&anno=2016

[71] A. Leone, L. Cancellieri, A. Guerriero, A. Cammelli, *The impact of the Bologna Process on the labour market: a national data warehouse regarding universities and graduates,* European University Information Systems, EUNIS, Santiago de Compostela (ES), 2009.

[72] A. Leone, L. Cancellieri, A. Guerriero, A. Cammelli, *Using Microsoft Analysis Service to analyze graduates' performances and working conditions*, European University Information Systems, EUNIS, Warsaw (PL), 2010.

[73] AlmaLaurea website, *Indagine 2014 Stella*, 2018, https://www.almalaurea.it/universita/occupazione/indagini_stella

[74] Apache Jena website, *Apache Jena Fuseki documentation*, 2018, https://jena.apache.org/documentation/fuseki2/

[75] D.M. 270/2004

[76] D.M. 509/1999

[77] J. Bertin, *Graphics and Graphic Information Processing*, Walter deGruyter, Berlin, 1981

[78] N. Shedroff, *Information Design*, Edited by R. Jacobson, MIT Press, Cambridge, 1999

[79] R. Spence, *Information visualization*, Vol. 1. , Addison-Wesley, New York, 2001.

[80] S. Card, J. D. Mackinlay, and B. Shneiderman, *Information visualization*, Human-computer interaction: Design issues, solutions, and applications, 2009

[81] J. Larkin and H. Simon. *Why a diagram is (sometimes) worth ten thousand words,* Cognitive science 11.1, 1987

[82] ClicLavoro, “Open Data: Motore dell’economia UE”, Newsletter num.2, February 2018