



Linked data and semantic web technologies to model context information for policy-making

Antonella Carbonaro¹

Received: 19 December 2018 / Accepted: 5 June 2019 / Published online: 11 June 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Currently, several datasets released in a Linked Open Data format are available at a national and international level, but the lack of shared strategies on the representation and meaning of knowledge related to the publishing community makes it difficult to compare and use them. The paper proposes the use of semantic technologies and linked open data in order to ensure standardized frameworks for the representation of concepts in policy-making. The low-level data can thus be transformed into an enriched information model that allows its reuse and a logical reasoning on the knowledge representation.

Keywords Linked Open Data · Semantic technologies · Open Government Data · RDFS · OWL

1 Introduction

Linked Open Data (LOD) principles have been adopted by an increasing number of data providers over the last years, leading to the creation of a global dataspace containing billions of assertions about geographic locations, people, companies, books, scientific publications, proteins, online communities, etc. Topics such as ontology building, use of semantics technologies and future applications that will be supported by these technologies are becoming important research areas in their own right (Ristoski and Paulheim 2016). As the main data publisher, governments without a doubt plays an important role to realize such value as the governments has been collecting and keeping an enormous amount of public data underutilized for long. This has led the term Open Government Data (OGD) become a prominence recently among researchers. Level of understanding of what is open government data among government agency also plays an important role in determining the adoption of open government data initiatives. Governments around the world still have a long way to go in terms of fulfilling what many see to be the great promise of open data. In its last edition, the Open Data Barometer (ODB) covers 115 countries and jurisdictions, a 25% increase on coverage from the

last edition (Open Data Barometer 2017). The findings from the ODB show that while some governments are advancing towards these aims, open data remains the exception, not the rule. In fact, only 7% of the data is fully open, only one of every two datasets is machine readable and only one in four datasets has an open licence. While more data has become available in a machine readable format and under an open licence since the first edition of the Barometer, the number of global truly open datasets remains at a standstill. Open innovation has emerged as an important concept in both academic research and industrial practice, and it is now also becoming increasingly important in the public policy domain. Open data has significant potential to foster innovation. For example, innovation and significant economic and social value can be created by using datasets such as map data, public transport timetables, statistical data and data on international trade or crime. Open data has the potential to make key public services—such as health, education and environmental management—more effective and inclusive. In the area of education, LOD has been used to share data such as courses offered by universities, national and international statistical data, organizational data and educational resources (Pereira et al. 2017). There are government datasets which contribute significantly to LOD related to education by, for example, generating statistics about the records and results of educational institutions.

In this paper, we report on the experiences we made when using linked data and semantic technologies to model context information in policy-making. We describe three case

✉ Antonella Carbonaro
antonella.carbonaro@unibo.it

¹ Department of Computer Science and Engineering,
University of Bologna, Bologna, Italy

studies: the first one models concepts related to the statistical domain of graduate surveys; the second experience models data on a student's entire school career from primary to secondary school and the third framework models the Great and General Council of the Single Chamber Parliament of the Republic of San Marino, the body responsible for the country's legislative power, within which laws, decrees and regulations are presented, discussed, amended, approved or rejected. We believe that these are important experiences for others who want to adopt a similar strategy for their data, as well as for the further development of OGD representation models in policy-making.

The integration of graduate surveys statistics in the LOD scope can make an important contribution to the information available on the status of graduates in Italy. The availability of exhaustive information about graduates' employment conditions supports other sample surveys deriving from National Institute of Statistics and enriches them with more specific data through the numerous variables. In a similar way, the survey on the profile of graduates further improves the picture, giving exhaustive and reliable information about the quality of the study experiences of the graduates, through data collected at the end of their student careers, which also provide a vision of subjective aspects such as personal satisfaction. This data can complete those deriving from the EUROSTAT surveys.

Similarly, a LOD system could help to develop improvement actions on the school system, highlighting its strengths and weaknesses, to plan subsequent targeted actions that can be monitored in the short and long-term results (European Data Portal 2018). The system could help to answer the following type of questions. How much does the learning/teaching path of a school order affect a student's school career? To what extent can the discontinuity in the transition between school orders be considered physiological and when does it become pathological? How much does this discontinuity affect early school leaving? How much does this discontinuity affect early school leaving? Finally, what can schools do to reduce the negative effects of transition and implement more targeted and effective guidance actions?

An example that reiterates the importance of the use of open data and semantic technologies is the lack of data updating in OGD. In fact, it is common to find discordant and/or outdated data on the web. You can easily realize this by browsing some pages of Wikipedia also related to the government context. Through the semantic representation, it will be possible to refer to the laws or the compositions of the single organisms exploiting the URIs, to represent in univocal way eventual changes or modifications.

The systems presented are able to: (1) establish a common semantics in describing the domain information; (2) create a LOD infrastructure; (3) support a more general management and sharing of knowledge; (4) provide semantic reasoning.

The novelty of the proposed approach lies in exploiting SW technologies to explicitly describe the meaning of the domain concepts and to facilitate interoperability and data integration, in order to construct a unified interlinked data model and enable semantic reasoning capabilities over it. A number of evaluations found within proposed approaches are discussed, and from them we extract challenges that hinder OGD initiatives from reaching their full potential. In compliance with the open paradigm, the structured data proposed are freely available and unconstrained in proprietary applications.

An up-to-date analysis of related works is presented in Sect. 2. Section 3 defines the main requirements that need to be taken into account when modelling context information for policy-making. Section 4 describes how linked data and semantic web technologies can be used to model, share and interlink context information. Section 5 proposes the use of open semantic approach to graduate's survey domain. Section 6 reports our approach to increase continuity by decreasing distances in the school. Section 7 describes governmental RDF/OWL common data model to represent the OGD entities. Finally, some considerations and conclusions close the paper.

2 Related works

Whilst government organizations have begun to enhance transparency by communicating and interacting with citizens via the Web, the development of appropriate Web of Data strategies has demanded a better understanding of user requirements for tailoring solutions. A major proportion of these open data consists of statistics, such as financial and social indicators; for example, the vast majority of datasets published on the open data portal of the European Commission (<http://open-data.europa.eu>) are provided by Eurostat and thus are of statistical nature. Many Italian public institutions have provided their LOD and the Italian government has proposed centralized management of them (<http://www.datiopen.it/>). In particular, the Italian government has taken several actions through the Digital Italian agency. These include, among others, the publication of national guidelines for the valorization of the public sector information, the definition of a license and the creation of a centralized catalogue of the open data of public administration (<http://dati.gov.it>). Despite this centralization, in some domains there are not yet standards to facilitate interoperability and data integration. International and national governments, for example, lack shared strategies concerning the definition of concepts related to the statistical publishing community, in particular statistics on graduates (McBride et al. 2018).

We can think of a statistical dataset as a multi-dimensional space or hyper-cube, indexed by dimensions that

define what the observation applies to along with metadata describing what has been measured, how it was measured and how the observations are expressed. In the case of cubes, LOD could enable the easy discovery and integration of multiple cubes on the Web and thus perform analytics on top of integrated but previously isolated cubes (Kalampokis et al. 2013). A fundamental step towards this vision is data cube (QB) vocabulary, which enables such information to be represented using the RDF standard and published following the principles of LOD (Cyganiak and Reynolds 2018). Recently, some tools have been developed to model data cubes according to the QB vocabulary; however, they present some limitations regarding the functionalities they provide, as well as their capability to be used in complex scenarios in an integrated manner (Kalampokis et al. 2014). Salas et al. (2012) propose two tools that use QB vocabulary to provide the representation of statistical data. OLAP2DataCube and CSV2DataCube allow both the analysis of a large amount of data and statistical data available in CSV files spreadsheets, and their transformation into RDF.

Within the SW context, ontologies play a central role in data representation, since they explicitly define concepts and relationships related to a particular domain in a structured and formal way (i.e., ontologies are machine-processable) (Reda et al. 2018; Carbonaro and Ferrini 2007; Carbonaro 2010a, b). We are able to transform raw data that are produced on a large scale by humans and machines into knowledge capable of supporting smart decision-making, innovative services, new business models and innovation (Lytras et al. 2018). Scheider et al., highlight the different roles that semantic technology may have in developing representation and reuse of data analysis across communities of practice, and investigate the role of SW in current analysis and workflow tools (Scheider et al. 2017).

The state of the art of open data availability in the education domain is currently quite fragmented (Aslam and Aljohani 2018). Taking the portal of the LOD of the Italian public administration as a reference point, researching the topic “degree” returns only datasets relating to a 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 (<https://www.europeandataportal.eu/>), which collects data from single national sources and merges them in a bottom-up modality, guaranteeing standardization by observing the principles of the open data paradigm. Despite the difficulties of attaining a complete, holistic picture, the LOD phenomenon is growing with regard to the theme of education and contributes to creating global knowledge which is very important for future generations (<https://www.europeandataportal.eu/highlights/open-data-schools>).

Ever more governments around the world are defining and implementing OGD strategies, there is still research

to be done to enable automated and scalable assessment as well as comparison of open data portal quality (Kubler et al. 2018). One of the challenges for properly comparing data portals is the development of multiple quality indicators, covering the different aspects of open data in e-government, for example, retrievability of the metadata and data and accuracy in the description of underlying resources. This is even more challenging because there exist several portal software frameworks, leading to a non-uniform publication of open data sets. Many datasets do not provide standardized description fields for geospatial and temporal properties about the datasets’ content or have free form fields to specify the format, license, keywords or descriptions. This results in only partially machine understandable descriptions, leading to the challenge of mapping the terms to known concept hierarchies such as DBpedia or WikiData. Metrics that require to inspect the content of a dataset and metrics that require a manual assessment are currently out of scope of this study. Automated and scalable assessment is an open research perspective.

3 Characteristics of context modelling systems

The main requirements that need to be taken into account when modelling context information for policy-making are:

3.1 Heterogeneity

Context information models have to deal with a large variety of context information sources that differ in their update rate and their semantic level. A context model should be able to express different types of context information and the context management system should provide management of the information depending on its type. For example, information provided by the user (like user profiles), in general does not need additional interpretation. Context data obtained from dataset or digital libraries (such as the British Museum, which has made their collection available as linked data, representing more than 100 million triples, or the Bibliothèque Nationale de France, which made available information about 30,000 books and 10,000 authors in RDF, representing around 2 million triples) is often static. While some context information derived from sensors provide rather raw data (for example Linked Sensor Data [http://wiki.knoesis.org/index.php/SSW_Datasets] have been introduced as an application of the linked data principles to observation data) that has to be interpreted before being usable by applications (Reda et al. 2018; Carbonaro and Santandrea 2018).

3.2 Relationships and dependencies

One of the most important characteristics of context modelling systems is relationships between context information entities. For example, we want to express logical constraints and detailed relationships such as disjoint, inverse, part-of, and so on. All the work involved in relationship and dependencies representation can directly benefit learners by helping them to visualize and comprehend the relationships between concepts in their domain, as understood by more experienced practitioners. This can trigger associative ways of processing, reflecting and analyzing information.

3.3 Imperfection

Due to its dynamic and heterogeneous nature, educational context information may be of variable quality. For example, the context information may be incomplete or conflicting with other context information. Thus, a good context modelling approach must include modelling of context information quality to support reasoning about context. Reasoning can also be used for automatically detecting inconsistency of the knowledge base.

3.4 Reasoning

It is important that the context modelling techniques are able to support both consistency verification of the model and context reasoning techniques. The later can be used to derive new context facts from existing context facts and/or reason about high-level context abstractions that model real world situations. The intrinsic potential of context representation can be exploited using sophisticated data analysis techniques such as automatic reasoning to find patterns and extract information and knowledge in order to enhance decision-making and deliver better educational resources and feedback to users.

We believe that linked data, ontologies and reasoning technologies can be considered a natural extension to context modelling approaches to meet the needs of heterogeneity, imperfection and need for reasoning. Using these approaches can improve transparency, foster innovation by exploiting the social and economic value of published data and foster the active participation of citizens in governance processes. The above motivations, while not being the sole ones, are the foundations for most open government data initiatives.

4 Linked data and semantic web technologies to model context information

The term Linked Data (LD) refers to a set of best practises for sharing and interlinking structured data and knowledge on the Internet by using standard web technologies (Bizer et al. 2011). The primary goal of the LD initiative is to make the Web not only useful for publishing documents, but also for sharing and interlinking single pieces of data. The movement is driven by the idea that the SW technologies facilitating the data sharing, integration, and analysis on a global scale could revolutionise the way we manage knowledge just like the Web revolutionised information sharing and communication over the last two decades.

Technologically, the core idea of LD is to use the Internationalised Resource Identifiers (IRIs) (Ishida 2008) for univocally identifying arbitrary entities and concepts. Information about entities referred by IRIs can be simply retrieved by dereferencing the IRI over the HTTP protocol.

Data about entities and concepts are then represented through the Resource Description Framework (RDF) (Cyganiak et al. 2014) language. RDF is a standardized data model which uses graphs to represent information and facts by means of triples in the form subject, predicate, object. Whenever a Web client resolves an IRIs associated to a triple's subject of a resource, the corresponding web server provides an RDF description of the identified entity, these descriptions can contain links to other RDF graphs in the triple's object. Whenever an application resolves a predicate IRI, the corresponding server responds with a RDF Schema (RDFS) (Brickley and Guha 2014) or Web Ontology Language (OWL 2012) definition of the link type, that is a vocabulary or an ontology. Ontologies are a key aspect of the SW since they enable interoperability among different systems by providing an agreed-upon terminology such as the basic terms and relations in a domain of interest, and as well as rules how to combine these terms.

Because LD is based on standards for the identification, retrieval, and representation of information and knowledge, and scattered entities are interconnected by links, it is possible to crawl the entire data space, fuse data from different sources, and provide expressive query capabilities over aggregated data, similarly to how a local database is queried today. For this purpose, the Simple Protocol and RDF Query Language (SPARQL 2013) is the standard language for querying, combining and consuming structured data in a similar way SQL does this by accessing tables in relational databases. Since LD is exclusively based on open web standards, data consumers and domain experts can use generic tools to access, analyze, and visualize data.

Moreover, LD make use of ontologies to formally define the meaning of entities and resource so that they do not limit the ability of machines to process data automatically.

Semantic Web (SW) describes a new way to make web content more meaningful to machines. The SW architecture is based on a layered approach, and each layer provides a set of specific functionalities. Semantic layers, on the top of the stack, include ontology languages, rule languages, query languages, logic, reasoning mechanisms, and trust. Ontologies, as a source of formally defined terms, play an important role within knowledge-intensive contexts such as the one described in this article. Ontologies can be reused, shared, and integrated across applications, and aim at capturing domain knowledge in a generic fashion and provide a common agreed understanding of the domain. Ontologies constitute the backbone of the SW expressing concepts and relationships of a given domain, and specify complex constraints on the types of resources and their properties.

Rule languages allow writing inference rules in a standard way which can be used for automatic reasoning. Among several standards of rule languages, there are RuleML and SWRL (Horrocks et al. 2004). The latter combines RuleML and OWL, and includes a high-level abstract syntax for Horn-like rules. SW technologies are a promising way for the integration and exploitation of food, nutrition, activity and personal data. In this context, ontologies enable the formal representation of the entities and their relationships and the associated background knowledge. On the highest layers there are logic and reasoning, logic provides the theoretical underpinning required for reasoning and deduction. First order logic and description logic (DL) (Baader et al. 2003) are frequently used to support the reasoning system which can make inferences and extract new insights based on the resource content rely on one or more ontologies.

Reasoning is the process of extracting new knowledge from an ontology and its instance base and represents one of the most powerful features of SW technologies especially for dynamic and heterogeneous environments. A semantic reasoner is a software system whose primary goal is to infer knowledge which is implicitly stated by reasoning upon the knowledge explicitly stated, according to the rules that have been defined. Reasoners are also used to validate the ontology, that is, they check its consistency, satisfiability and classification of its concepts to make sure that the ontology does not contain any inconsistencies among its term definitions.

SWRL language can enhance the ontology language by allowing us to describe relations that cannot be described using DL. An SWRL rule is defined in form of if-then clauses containing logical functions and operations. Additionally, SWRL provides many sets of built-in functionality, such as mathematical functions and string operations. This scenario promotes interoperable communication among various information

technology systems and can be used for automatic inferencing/reasoning, and to check-out logical data consistency.

The conceptual architecture of such a system is illustrated in Fig. 1 comprising of four components namely data retrieving layer, data processing layer, service layer, and presentation layer. Data retrieving layer collects domain datasets from users or automatically from remote servers. Data processing layer transforms input data in semi-structured formats into an RDF graph, datasets are thus semantically annotated according to reference ontologies and stored within a triple store server. Service layer controls data access and bridges the clients to the system via service protocols. Presentation layer allows users to interact with the system using either the web based access or the SPARQL endpoint.

Data Processing Layer is the core of the architecture and represents the domain knowledge base, which updates the semantically rich annotated structured data. The repository maintains different types of knowledge statements: Tbox, that is terminological knowledge that defines classes, properties and relationships, and Abox, that is assertional knowledge to represent the individual instances of the concepts and defines them with factual statements. Individual instances are marked up with ontologies to discover the attributes and class membership and relationship of different entities. The domain knowledge includes the identification of the key concepts used or referred to in the domain and identification of different vocabularies or metadata standards that can be used to further enrich the annotation of key concepts to provide a broader contextualization of the concept. Personalized Layer is where personalization could be developed by matching resources with personal profile through knowledge base querying mechanism using SPARQL.

Although the number of initiatives seeking to publicly disclose government data is dramatically increasing, it is still a major challenge to reach the full potential of OGD and to support all stakeholders in the publication and consumption of this data. The heterogeneous nature of data formats and data structures used by public administrations remains one of the major obstacles contributing to this challenge. These heterogeneities are a technical barrier to both data producers and data consumers. We propose the use of semantic technologies and linked open data in order to ensure standardized frameworks for the representation of concepts in policy-making and to make possible the comparison and aggregate analysis of government data. This interconnection of the variety of publicly available data sources can significantly facilitate reuse, exploitation, and possible extension of data.

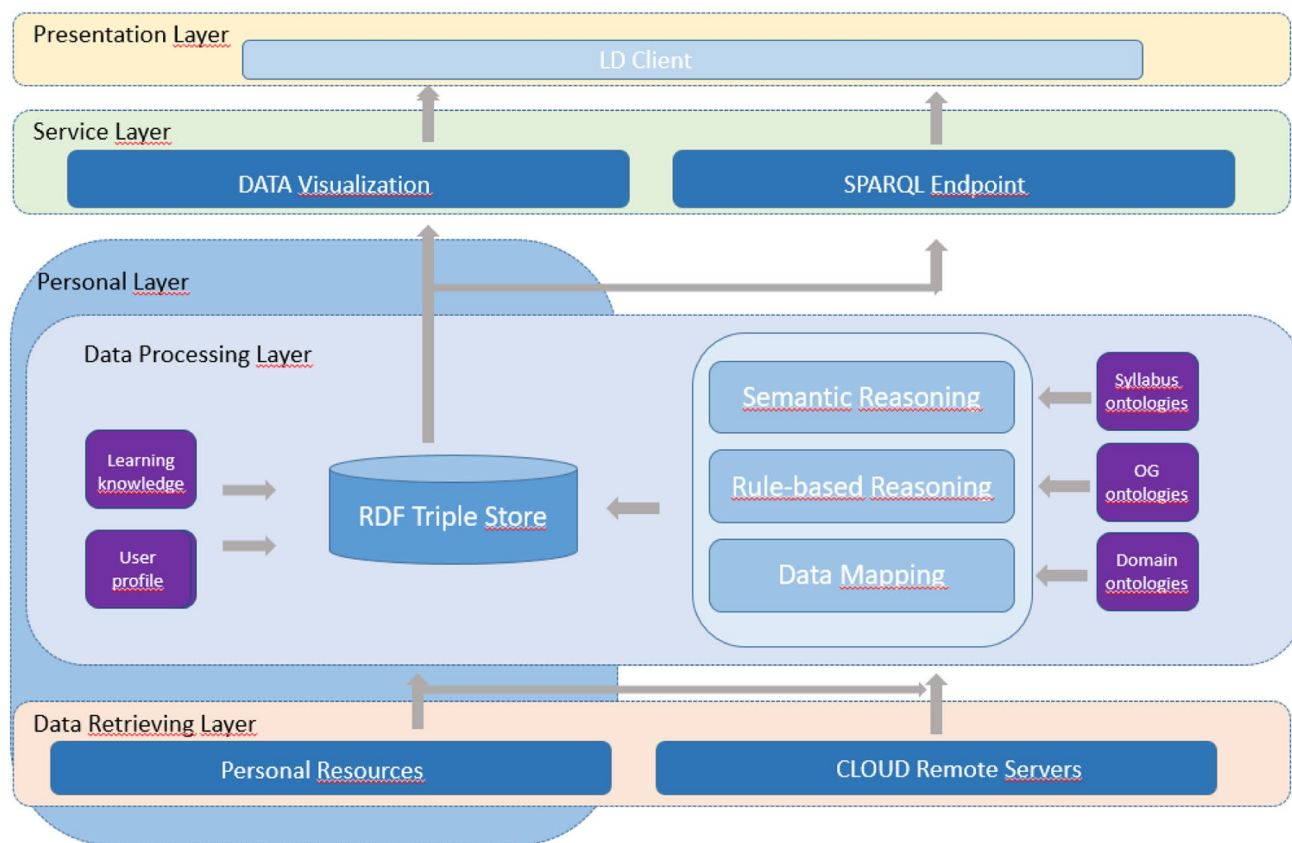


Fig. 1 Conceptual architecture with four layers from data retrieving layer to presentation layer

5 Graduate's survey domain

The ontology-based system for the graduate's survey domain transforms low-level data into an enriched information model encoded using RDF describing the different peculiarities of the graduates. A significant contribution to graduate statistics in the national sphere is the work done by the AlmaLaurea inter-university consortium (<http://www.almalaura.it/en>). Founded in 1994 after an initial project begun by the Statistical Observatory of the University of Bologna, the consortium—supported by the Italian Ministry of Education, Universities and Research—has as its main mission the production of statistical surveys about the situation of the Italian graduates (Leone et al. 2010). These surveys are widely representative of the sector due to the high number of member universities (75 by early 2018), which guarantees a coverage of more than 90% of Italian graduates. This diffusion makes the AlmaLaurea surveys a reference point for the academic community and the economic and political worlds. The high number of questionnaires (more than 200,000 per year) produces a significant dataset (AlmaLaurea 2018). We developed the Semantic Web for AlmaLaurea (SW4AL), an ontology-based system for the graduate surveys domain which transforms low-level data into an

enriched information model encoded using RDF datasets describing the different attributes of graduates. SW4AL uses standard Semantic Web (SW) technologies promoted by W3C. This effort is useful for the expression of the AlmaLaurea data as meaningful information, providing their possible reuse in the open data scope via the publication of the resulting RDF datasets. Indeed, ontology-based SW technologies can support interoperability among data and can be used for semantic annotation, information sharing, resource discovery and knowledge transformation (Bischof et al. 2018; Carbonaro and Ferrini 2008; Riccucci et al. 2007; Carbonaro 2010a, b). In so doing, the data becomes usable in different contexts, for instance as the basis of mashup applications; it also makes it possible to enhance their value by combining them with external sources, getting the rid of the information silos problem. For instance, integration with international open datasets, like the ones available through the European Union open data portal, can help to compare Italian graduates' performances with those from other countries. The SW4AL ontology developed actually consists of 74 classes, 121 object properties, 107 data properties and 1628 axioms. The dataset contains 3,161,153 triples. Furthermore, the system offers ontology-based data access to end-users. They could pose relatively complex and ad hoc

queries in an easy and intuitive way, extracting, combining and interpreting data in previously unforeseen ways (Fig. 2). More details can be found in (Carbonaro and Santandrea 2018).

The evaluation of performance can be analysed through two different aspects: the time needed for the generation of the triple store and the throughput of the data visualization tools. The first aspect depends strongly on the computational power of the SQL server database machine; as the extraction occurs once a year, it does not represent a major concern. Regarding data visualization tools, most of the responsibilities for performance are prerogatives of the 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 has been noticed thanks to the usage of functional constructs of JavaScript. A satisfying benchmark is nonetheless guaranteed with the generation of all the survey data for a 3-year period (circa three million triples).

The Graduate's survey ontology represents characteristics and performances of the graduates considering criteria such as 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. Employment condition ontology represents the insertion of the graduates in the business world by collecting data deriving from interviews conducted at 1, 3 and 5 years from the achievement of the bachelor/master 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. For example, graduate's survey ontology comprises Profile class (its instances are the subjects of the triples regarding the profile survey), Employment class (to represent the employment condition statistics), Course class (to represent 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), University class (to describe the university institution, it can be defined by other ontologies) and Degree class (to represent horizontal division about the hierarchical system).

Possible scenarios of the usage of these data can be hypothesized according to the existing open datasets regarding the targeted job placement of the graduates, like the open data released by the Italian Ministry of Labour and Social Policy (<http://dati.lavoro.gov.it/Pages/home.aspx>). Moreover, the integration with international open datasets like the ones exposed by the European Union open data portal (<https://open-data.europa.eu/>), can help to compare graduates' performances with those from others countries, process which leads to an increase of the knowledge in the domain.

Reasoning on this enriched information can lead to interesting considerations and conclusions. 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 and location. This last one can help the institutions and the universities to verify the causes of different performances, with possible significant reflexives on the data knowledge and on the economic and social growth of the country.

6 Decrease distances to increase continuity

Numerous national and international researches have focused on the analysis of students' evaluations (Carbonaro and Ravaioli 2017; Carbonaro 2012). From the OCSE-PISA (Programme for International Student Assessment) surveys, which since 2003 have involved an increasing number of countries in the world, to the INVALSI national tests, proposed to schools since 2004, to the research produced by the Agnelli Foundation, which led to a substantial report published in 2014 and to the creation of the site *eduscopio.it*, on which students can search for the results of schools, to find those that best prepare for university studies.

Compared to existing research, *RiminInRete* (<http://www.rimininrete.net>) has the following peculiarities: (1) it is a project that starts from schools: schools are no longer just the subject of research, but become conscious protagonists of the process of improvement, of which the collection of data is the essential basis. (2) It is a networked project that starts from the awareness that a territory grows and improves by joining forces and producing systemic actions. (3) It is a project that foresees to go down from macro to micro analysis in which each school will be able to study its own data with reference to some specific criticalities on which it intends to focus in its own improvement plan.

Therefore, the aim of the project is not simply a goal of customer satisfaction, but the activation of concrete and contextualized actions, for the overall improvement of the training offer, in respect of the proactive action that must be the basis of the organizational and didactic activities of the school.

The project represents data on, for example, the results of leaving secondary school in the first degree, comparing them with the results of entering secondary school in the second degree. It has been hypothesized that the effects of discontinuity in this transition and the results of a more or less effective school orientation can have significant consequences for both on the future regularity of the course (thus also determining phenomena of dispersion), and more generally on the relationship of students with the school (motivation to study, psychological and relational difficulties increasing


<p>Region selection</p> 	<p>Survey selection</p> <p>Year of execution: 2016</p> <p>University: University of BOLOGNA</p> <p>Degree type: Bachelor's degree</p> <p>Course: astronomy (BOLOGNA)</p> <p><input type="checkbox"/> Aggregate to D.M. 270 courses the relative D.M. 509</p>	<p>Question selection</p> <p><input checked="" type="checkbox"/> Currently employed graduates</p> <p><input type="checkbox"/> Graduates currently enrolled to a master's degree course</p> <p><input type="checkbox"/> Graduates not working and not looking for a job but which attend a university course or internship or apprenticeship</p> <p><input type="checkbox"/> Graduates using in a high measure the competences acquired with the degree</p> <p><input type="checkbox"/> Net monthly salary (average)</p> <p><input type="checkbox"/> Average satisfaction for the current job</p>
<p>Selected criteria</p> <p>Year: 2016</p> <p>University: University of BOLOGNA</p> <p>Degree type: Bachelor's degree</p> <p>Course: astronomia (BOLOGNA)</p> <p>Selected questions: 1</p>	<p>Query text</p> <pre> prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> prefix owl: <http://www.w3.org/2002/07/owl#> PREFIX alma: <http://www.almaurea.it/opendata/ontologies/default#> PREFIX profilo: <https://www.almaurea.it/opendata/ontologies/profile#> PREFIX occ: <http://www.almaurea.it/opendata/ontologies/employment#> </pre>	<p>Select survey type (years from graduation)</p> <p><input type="radio"/> 1 year</p> <p><input checked="" type="radio"/> 3 years</p> <p><input type="radio"/> 5 years</p>

Fig. 2 Wizard form for the employment condition queries. In order to guarantee the full utilization of the open knowledge base generated, we developed a wizard interface for the incremental creation of SPARQL queries

in recent years, etc.). The students currently involved in the project are more than 50,000 and the schools managed more than 160.

For example, the following Fig. 3 shows the histogram of the students according to the schools of origin shown on the map. It also shows the averages with which students enter the school year under examination and how students are divided within the school, depending on the field of study they have chosen. For each page, you can filter the data by school year, school, class and section. In this way, you can focus on more specific data related to the context of interest.

7 Governmental RDF/OWL common data model

In the context of OGD, semantic web technologies such as RDFS and OWL can be used to represent the entities that compose the Grand and General Council, the unicameral Parliament of the Republic of San Marino. Inspired by the latest implementations in this area, like the ontologies made available by the English Parliament or the European Legislation Identifier, we wanted to describe the structure and duties of each component of the Parliament, including proceedings like laws and decrees.

The ontology is made of nine macro-ontologies, each of them covers a different aspect: in fact, it also models the related institutions that have to do with the Parliament, like the Congress of State and the Captains Regent (the Heads of State). RDFS vocabularies and owl ontologies comprise: (1) Base vocabulary: provides a basic model of people, properties, institutions; (2) Grand and General Council vocabulary: an ontology to model the components and tasks assigned to the Grand and General Council; (3) Commissions vocabulary: model for the parliamentary commissions; (4) Congress of State vocabulary: classes and properties regarding the Congress of State, the Secretariates of State and the Ministers; (5) Council of 12 vocabulary: description of the Council of Twelves; (6) Councilial Works vocabulary: basic model for the councilial works: convocations, sessions and agenda; (7) International Delegations vocabulary: a model for the international parliamentary delegations of the GGC, e.g. European Council and OCSE; (8) Regency vocabulary: entities modelling the Regency of the Republic; (9) Textual Products vocabulary: classes and properties about the textual products of the GGC, like norms or documents.

The long-term objective is to foster the implementation of open data platforms for the institutions of San Marino, a country that is moving forward at a fast pace, but where certain types of data are not yet publicly available or, if available, they're not in a machine-readable format. This ontology could help to improve and simplify the works of the legislative power of the Republic by building applications

on top of it. Both RDFS and OWL artifacts are available at <https://nicorsm.github.io/cgg-ontology/>.

8 Considerations and conclusions

There are many situations where it would be useful to be able to publish multi-dimensional data in such a way that they can be linked to related datasets and concepts. OGD are one of the most important sources of information, relevant to large numbers of domains from government to business and education. We believe that context information data are a foundation for policy prediction, planning and adjustments, and underpin many of the mash-ups and visualizations we see on the Web. The shared aims of the data publishing community mean that its members could benefit from SW technologies. Those aims include making data easy to locate and open up for reuse, particularly for analysis and visualization. For example, this paper proposes how to use statistical datasets linked into the wider Web of Data to improve the process of filtering and querying statistical data. These observations often reference the geographical area to which the statistic applies and some extra analysis can be made possible by resolving information about that area, such as size of the territory and civilization degree. Our research demonstrates the advantages of explicitly encoding specific domain concepts by formally representing the semantics of the collected data, the domain and their relationships. Specifically, in this paper we proposed three LOD ontology-based systems which transforms low-level data into an enriched information model encoded using RDF describing the different attributes. We propose to use SW technologies to map information onto a specialized domain model by providing support for logical reasoning. The presented approach can significantly facilitate sharing, exploitation and creative reuse of existing data sources, thus fully exploiting their intrinsic potential. We suggest using this approach wherever there is the need to create knowledge from information and information from data using LOD, semantic representation, reasoning technologies and incorporating domain knowledge into the computation. Although the benefits of open data are undeniable, one aspect that still deserves much attention is public participation in open public data initiatives, which needs to be improved. To encourage participation from external stakeholders we believe that it may be important to provide examples of successful use case studies. This paper goes in this direction by providing three use cases that can be taken as a starting point for the implementation of new services promoting the innovative potential of developers and other stakeholders. Moreover, the development of these projects allowed us to interact with different stakeholders and to identify their needs. The need for open, comparable and sharable data emerges clearly; often it is also clear

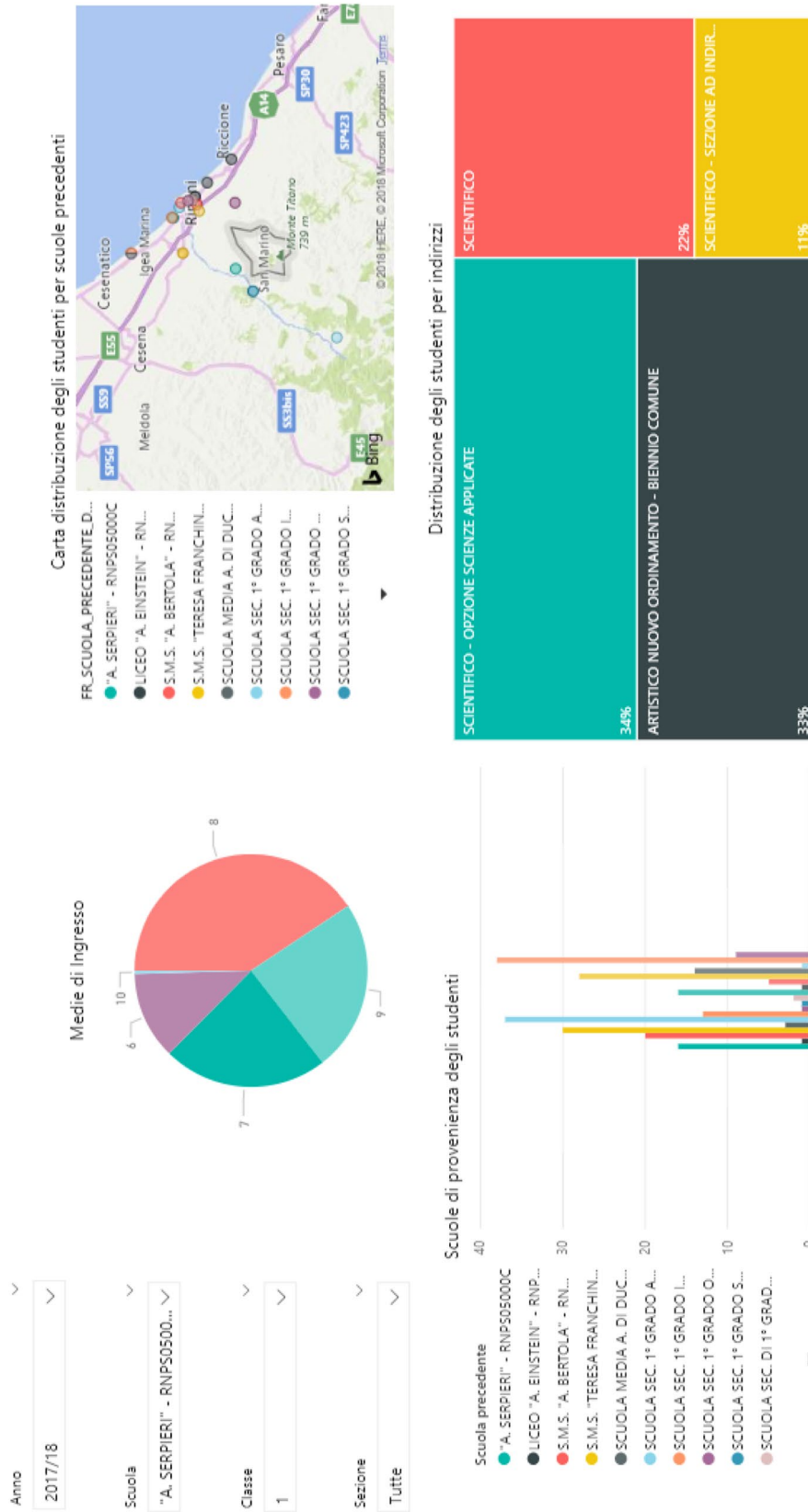


Fig. 3 Representation of the distribution of students for schools of source

that there is a need to have use cases to understand how to actually use them. The collection of examples and best practices of community organizations and other stakeholders use of data and the building of semantic data infrastructures would provide clear paths for communities to follow. Possible future developments of the work concern the integration of existing open datasets based on the functionalities that will be integrated. For example, with regard to targeted job placement of graduates, open data issued by the Ministry of Labour and Social Policy could be integrated. This opportunity goes towards the direction indicated by the same ministry, which through its job portal *ClicLavoro* has promoted open data as an “engine of the European Union’s economy” (<https://www.cliclavoro.gov.it>). Moreover, integration with international open datasets like the ones available through the European Union open data portal can help to compare Italian graduates’ performances with those from other countries, a process which will lead to an increase of knowledge in the domain by providing a simple benchmark. In this way, reasoning on the enriched information can lead to interesting considerations and conclusions. One example of this might be the extraction of a time series that reports the different performances of graduates in given courses over several years. Another example might be a comparison of universities’ results based on their size and location. This case might help institutions and universities to verify the causes of different performances, with possible significant reflection on data knowledge and on the economic and social growth of the country.

Acknowledgements The author would like to thank Nicola Giancetti for the implementation and publication of the RDFS and OWL ontologies described in <https://nicorsm.github.io/cgg-ontology/>.

References

- AlmaLaurea (2018) Indagini e ricerche. <http://www.almalaurea.it/universita/statistiche>. Accessed 12 Mar 2018
- Aslam MA, Aljohani NR (2018) SPedia: a central hub for the linked open data of scientific publications. *IJSWIS* 13.1(2017):128–147 (Web)
- Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (2003) *The description logic handbook: theory, implementation and applications*. Cambridge University, Cambridge
- Bischof S, Harth A, Kämpgen B, Polleres A, Schneider P (2018) Enriching integrated statistical open city data by combining equational knowledge and missing value imputation. *J Web Semant* 48:22–47
- Bizer C, Heath T, Berners-Lee T (2011) *Linked data: the story so far. Semantic services, interoperability and web applications: emerging concepts*. IGI Glob 2011:205–227
- Brickley D, Guha RV (2014) RDF schema—W3C recommendation. <https://www.w3.org/TR/rdf-schema/>. Accessed 16 Aug 2018
- Carbonaro A (2010a) Improving web search and navigation using summarization process. *Commun Comput Inf Sci* 111(PART 1):131–138
- Carbonaro A (2010b) WordNet-based summarization to enhance learning interaction tutoring. *J e-Learn Knowl Soc* 6(2):67–74
- Carbonaro A (2012) Interlinking e-learning resources and the web of data for improving student experience. *J e-Learn Knowl Soc* 8(2):33–44
- Carbonaro A, Ferrini R (2007) Ontology-based video annotation in multimedia entertainment. In: *Consumer communications and networking conference, 2007. 4th IEEE. Citeseer*, pp 1087–1091
- Carbonaro A, Ferrini R (2008) Personalized information retrieval in a semantic-based learning environment, social information retrieval systems: emerging technologies and applications for searching the web effectively, pp 270–288
- Carbonaro A, Ravaioli M (2017) Peer assessment to promote deep learning and to reduce a gender gap in the traditional introductory programming course. *J e-Learn Knowl Soc* 3:13
- Carbonaro A, Santandrea L (2018) A general semantic web approach for data analysis on graduates statistics. In: *IEEE conference of open innovation association, FRUCT*, pp 99–104
- Cygniak R, Reynolds D (2018) The RDF data cube vocabulary. <https://www.w3.org/TR/vocab-data-cube/>. Accessed 9 June 2018
- Cygniak R, Wood D, Lanthaler M (2014) RDF 1.1 concepts and abstract syntax—W3C recommendation. <https://www.w3.org/TR/rdf11-concepts/>. Accessed 16 Aug 2018
- European Data Portal, Education: Open Data in Schools (2018) <https://www.europeandataportal.eu/highlights/open-data-schools>. Accessed 12 Aug 2018
- Horrocks I, Patel-Schneider PF, Boley H, Tabet S, Grosz B, Dean M et al (2004) SWRL: a semantic web rule language combining OWL and RuleML. W3C Member submission 21, p 79
- Ishida R (2008) An introduction to multilingual web addresses. <https://www.w3.org/International/articles/idn-and-iri/>. Accessed 16 Aug 2018
- Kalampokis E, Tambouris E, Tarabanis K (2013) Linked open government data analytics. In: Wimmer MA, Janssen M, Scholl HJ (eds) *EGOV2013, LNCS, 8074*. Springer, 2013, pp 99–110
- Kalampokis E, Karamanou A, Nikolov A, Haase P, Cygniak R, Roberts B, Hermans P, Tambouris E, Tarabanis K (2014) Creating and utilizing linked open statistical data for the development of advanced analytics services. In: *Proc. of the 2nd International Workshop on Semantic Statistics (Sem-Stats2014) in conjunction with the 13th International Semantic Web Conference (ISWC2014), CEUR-WS proceedings*
- Kubler S, Robert J, Neumaier S, Umbrich J, Le Traon Y (2018) Comparison of metadata quality in open data portals using the analytic hierarchy process. *Gov Inf Q Elsevier* 35(1):13–29
- Leone A, Cancellieri L, Guerriero A, Cammelli A (2010) Using micro-software analysis service to analyze graduates’ performances and working conditions, European University Information Systems, EUNIS, Warsaw (PL)
- Lytras MD, Raghavan V, Damiani Ernesto (2018) Big data and data analytics research: from metaphors to value space for collective wisdom in human decision making and smart machines. *IJSWIS* 13(1):1–10
- McBride K, Matheus R, Toots M, Kalvet T, Krimmer R (2018) The role of linked open statistical data in public service co-creation. In: *Proceedings of the 11th international conference on theory and practice of electronic governance (ICEGOV ‘18)*, Atreyi Kanakhalli, Adegboyega Ojo, and Delfina Soares (Eds.). ACM, New York, NY, USA, pp 679–681
- Open Data Barometer, 4th edition (2017) *Data World wide web foundation, datasets and report*, [online]. <http://opendatabarometer.org/4thedition/report/>. Accessed 10 June 2019
- Pereira CK, Siqueira S, Nunes BP, Dietze S (2017) Linked data in Education: a survey and a synthesis of actual research and future challenges. *IEEE Trans Learn Technol* 11:400–412

- Reda R, Piccinini F, Carbonaro A (2018) Towards consistent data representation in the IoT healthcare landscape. In: ACM DH'18: International Digital Health Conference, April 23–26, Lyon, France
- Riccucci S, Carbonaro A, Casadei G (2007) Knowledge acquisition in intelligent tutoring system: a data mining approach. In: Mexican International Conference on Artificial Intelligence. Springer, pp 1195–1205
- Ristoski P, Paulheim H (2016) Semantic web in data mining and knowledge discovery: a comprehensive survey, web semantics: science, services and agents on the world wide web, vol 36, pp 1–22
- Salas PER, Martin M, Da Mota FM, Auer S, Breitman K, Casanova MA (2012) Publishing statistical data on the web. In: International Conference on Semantic Computing, 6th ed, pp 285–292
- Scheider S, Ostermann FO, Adams B (2017) Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis. *Future Gener Comput Syst* 72:11–22
- W3C OWL Working Group (2012) OWL 2 Web Ontology Language—W3C Recommendation. <https://www.w3.org/TR/owl2-overview/>. Accessed 16 Aug 2018
- W3C OWL Working Group (2013) SPARQL 1.1 Overview—W3C Recommendation. <https://www.w3.org/TR/sparql11-overview/>. Accessed 16 Aug 2018

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.