ContrattiPubblici.org, a Semantic Knowledge Graph on Public Procurement Information

Giuseppe Futia[1], Federico Morando[2], Alessio Melandri[2], Lorenzo Canova[1], and Francesco Ruggiero[1]

[1] Nexa Center for Internet and Society,
Department of Control and Computer Engineering, Politecnico di Torino, Italy
https://nexa.polito.it
{giuseppe.futia,lorenzo.canova,francesco.ruggiero}@polito.it
[2] Synapta Srl, Italy
https://synapta.it/
{federico.morando,alessio.melandri}@synapta.it

Abstract. The Italian anti-corruption Act (law n. 190/2012) requires all public administrations to spread procurement information as open data. Each body is obliged to yearly release standardized XML files, on its public website, containing data that describes all issued public contracts. Though this information is currently available on a machine-readable format, the data is fragmented and published in different files on different websites, without a unified and human-readable view of the information. The ContrattiPubblici.org project aims at developing a semantic knowledge graph based on linked data principles in order to overcome the fragmentation of existent datasets, to allow easy analysis, and to enable the reuse of information. The objectives are to increase public awareness about public spending, to improve transparency on the public procurement chain, and to help companies to retrieve useful knowledge for their business activities.

Keywords: public procurement, linked data, knowledge graph

1 Introduction

In recent years the amount and variety of open data released by public bodies has been factually growing[3], simultaneously with the increase of political awareness on the topic[4]. Public Sector Information (PSI)[5], in the form of open

[3] See the Tracking the state of open government data report available at: http://index.okfn.org/. Last visited July 2016

[4] See national roadmaps and technical guidelines, as well the revised of the EU Directive on Public Sector Information reuse in 2013 guidelines

[5] Public Sector Information includes "any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording)" when produced by a public sector body within its mandate. See more details on Directive 2003/98/EC: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF

data, leads to a noticeable value for diverse actors and for different purposes, from transparency on public spending to useful knowledge for business activities. Open data is therefore a toolbox to improve relationships among governments, citizens and companies by directly enabling informed decisions. Nevertheless, access and reuse of data to build useful knowledge is extremely limited, mainly because of the fragmentation in different data sources and websites, which currently characterizes the publication of PSI.

As defined by a World Wide Web Consortium (W3C) issue proposed by Tim Berners Lee [2] on the publication of government data, linked data principles can be a modular and scalable solution to overthrow the fragmentation of information: "Linked data can be combined (mashed-up) with any other piece of linked data. For example, government data on health care expenditures for a given geographical area can be combined with other data about the characteristics of the population of that region in order to assess effectiveness of the government programs. No advance planning is required to integrate these data sources as long as they both use linked data standards." As stressed by Berners Lee, according to these precepts linked data serves to: 1) increase citizen awareness of government functions to enable greater accountability; 2) contribute valuable information about the world; and 3) enable the government, the country, and the world to function more efficiently.

Public procurement is an area of the PSI that could largely benefit from linked data technologies. As argued by Svátek[10], an interesting aspect of public contracts from the point of view of linked data is the fact that "they unify two different spheres: that of *public* needs and that of *commercial* offers. They thus represent an ideal meeting place for data models, methodologies and information sources that have been (often) independently designed within the two sectors." At the same time, linked data is beneficial in the public contracts domain since it gives ample space for applying diverse methods of data analytics, performing complex alignments of entities in a knowledge graph and developing data driven applications.

The contribution is structured as follows. Section 2 presents related works in the field of public procurement and spending information published according to linked data principles. Section 3 describes the Italian context and gives an overall view of public procurement data spread by public sector bodies. Section 4 illustrates the data processing pipeline to improve the quality of data source and to create the ContrattiPubblici.org knowledge graph. Section 5 shows results and potential use of the information structured in the graph. The last section describes conclusions and future advancements of the work.

2 Related Works

In this section, we report contributions in which public procurement and spending data is transformed and published as knowledge graphs, following the linked data principles.

Public procurement domain has already been addressed by several works and projects developed in the linked data field. One of the most notable is the LOD2 project, since it systematically addressed many phases of procurement linked data processing [10]. Such project exploits the Public Procurement Ontology PPROC[6].

There are several other initiatives: the TWC Data-Gov Corpus [4], Publicspending.gr [8], The FTS (Financial Transparency System) project [7], Linked Spending [6], LOTED [11] and MOLDEAS [1].

In particular, the TWC Data-Gov Corpus gathers linked government data on US financial transactions from the Data.gov project[7]. This project exploits a semantic-based approach, in order to incrementally generate data, supporting low-cost and extensible publishing processes, and adopting technologies to incrementally enhance such data via crowdsourcing. Publicspending.gr has the objective of interconnecting and visualizing Greek public expenditure with linked data to promote clarity and increase citizen awareness through easily-consumed visualization diagrams. The FTS (Financial Transparency System) project of the European Commission contains information about grants for EU projects starting from 2007 to 2011, and publishes such data according to the RDF[8] data model. Exploring such dataset, users are able to get an overview on EU funding, including data on beneficiaries as well as the amount and type of expenditure. Linked Spending is a project for the RDF conversion of data published by the OpenSpending.org, an open platform that releases public finance information from governments around the world. The project uses the RDF DataCube vocabulary[9] to model data in order to represent multidimensional statistical observations. LOTED[10] is focused on extracting data from procurement acts, aggregating it over a SPARQL[11] endpoint. Finally, MOLDEAS, between the other things, presents some methods to expand user queries to retrieve public procurement notices in the e-Procurement sector using linked open data.

3 Context and Data Source

The Italian Legislative Decree n. 33/2013 (DL33/2013) of March 14th, 2013[12] reordered obligations of disclosure, transparency and dissemination of information by public administrations. According to specific requirements defined by the decree (clause no.9 - DL33/2013), each body is required to create a specific section

[6] More details on Public Procurement Ontology PPROC available at: `http://contsem.unizar.es/def/sector-publico/pproc.html`

[7] Data.gov project website: `https://www.data.gov/`

[8] RDF (Resource Description Framework) is a standard model for data interchange on the Web. It represents a common format to achieve and create linked data

[9] DataCube vocabulary information: `https://www.w3.org/TR/vocab-data-cube/`

[10] LOTED project website: `http://www.loted.eu/`

[11] SPARQL (SPARQL Protocol and RDF Query Language) is a semantic query language for databases, able to retrieve and manipulate data stored in RDF format

[12] `http://www.decretotrasparenza.it/wp-content/uploads/2013/04/D.Lgs_.-n.-332013.pdf`. Last visit on July 2017

on its website called "Amministrazione Trasparente" (Transparent Administration). In this section, public administrations provide details related to public procurement, with particular emphasis on procedures for the award and execution of public works, services and supplies (clause no. 37 - DL33/2013). Such data is published on the basis of a precise XML Schema Definition[13] (XSD) provided by *ANAC - Autorità Nazionale Anticorruzione* (the Italian National Anti-Corruption Authority)[14], which has supervisory duties. After the publication on their websites, administrations transmit information on public contracts, in a digital format, to ANAC. ANAC then performs a preliminary check and releases an index file containing details related to the availability of data[15].

Public bodies can publish and transmit to ANAC two types of XML files. The first type contains the actual data on contracts until the publication date (January 31st of each year). As mentioned before, in order to facilitate the consistency of publications and the comparison of information, the structure of the document is defined by a precise XSD Schema[16]. The main structure of the XML file includes a section with the metadata of the dataset, reported in Figure 1, and a section containing multiple contracts. The metadata section lists diverse information including the first publication date and the last dataset update, the business name of the public body that spreads the dataset, the url of the dataset and the license. The section containing data on contracts includes: the identification code of the tender notice (CIG that stands for Codice Identificativo Gara), the description of the tender identified by the CIG, the procedure type for the selection of the contractor, the identification code (VAT number) and the business name of bidders (tender participants), the identification code and the business name of the tender notice beneficiary, the award amount, the amount paid, the date of commencement and completion of works (for more details of each field see Figure 2 and Figure 3). In Section 3.2 we describe the ontology used to map those fields in the linked data domain to build the semantic knowledge graph of ContrattiPubblici.org.

The second type of XML is an index that collects links to other XML files containing actual public procurement data (Figure 4)[17]. This data is available in machine-readable format according to a well-defined schema; a semantic layer to interconnect such data is necessary for the assessment of transparency, it is effective to raise awareness about public spending, and to provide useful information for enterprises.

[13] XSD is a W3C recommendation that specifies how to formally describe an XML document

[14] `http://www.anticorruzione.it/`

[15] The index is available at `https://dati.anticorruzione.it/#/l190`, clicking on the "Esporta" (Export) button

[16] The full representation of the XSD schema is available at `http://dati.anticorruzione.it/schema/datasetAppaltiL190.xsd`

[17] A more clear representation of the XSD schema is available at `http://dati.anticorruzione.it/schema/datasetIndiceAppaltiL190.xsd`

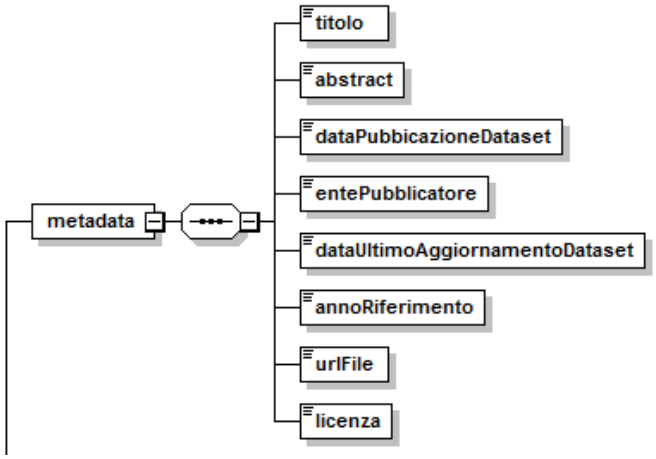

Fig. 1. XSD Schema of metadata

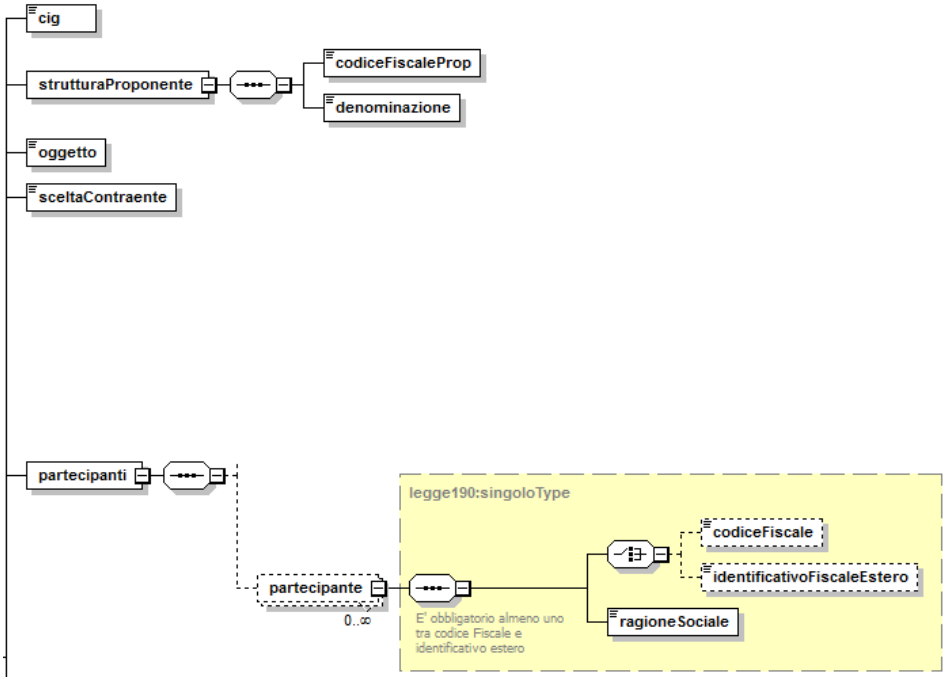

Fig. 2. XSD Schema of public procurement data - Part 1

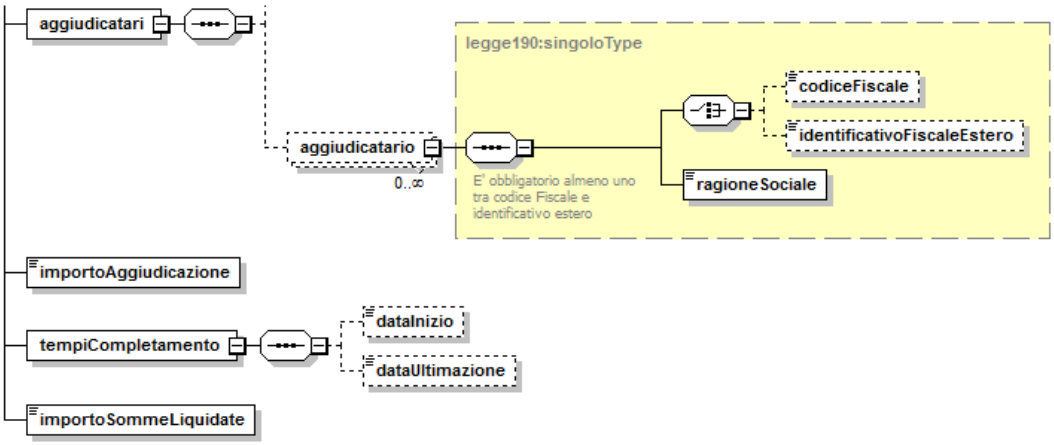

Fig. 3. XSD Schema of public procurement data - Part 2

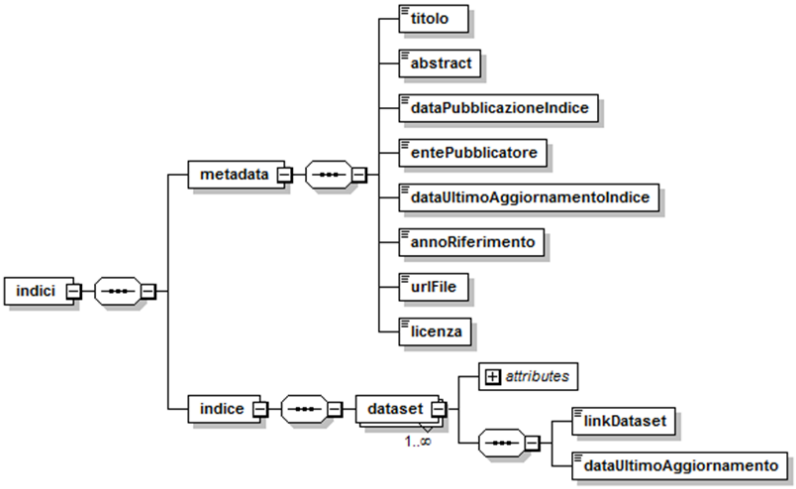

Fig. 4. XSD Schema of public procurement dataset index

3.1 Data Quality Issues

The quality of public contracts data is one of the most important issues to be tackled in order to reduce fragmentation and build a semantic knowledge graph. Let's consider, for example, a company that has participated in two calls for tender proposed by two different bodies. VAT number and business name of this company reported in the two XML files should be identical. However, some data errors may occur due to management processes and software: this prevents to generate a unique entity (the company itself) within the knowledge graph. For example, a VAT number can present a wrong character (*accuracy* issue), or even the field itself could be absent (*completeness* issue)[18]. Analyzing the VAT number issues, we have observed that 62,466 contracts (1.08% of the total) present accuracy problems like wrong characters, and 60,731 contracts (1.05% of the total) do not present the VAT number field in the data (completeness problems).

For this reasons, different checks must be implemented with the aim of correcting, where possible, the wrong data [12]. Section 4 describes the process we have implemented to tackle data quality issues in order to build a semantic graph upon public contracts data.

3.2 Ontology

In order to model the data source to build the ContrattiPubblici.org knowledge graph, we decided to use the Public Contracts Ontology (PCO) developed by the Czech OpenData.cz initiative[19]. The authors of this ontology are modeling "information which is available in existing systems on the Web" and "which will be usable for matching public contracts with potential suppliers" [5]. Therefore, the goal is to model a public contract as a whole, but without going into details of the public procurement domain.

In the PCO, a contract notice is a call for tenders, which may be submitted for the award of a public procurement contract. Therefore, we are able to map XML fields described in Section 3 into entities, classes and relations provided by the PCO. Figure 5 shows precisely the data model adopted for building the ContrattiPubblici.org knowledge graph. Although there is a significant degree of overlap between the XSD that describes the data model of Italian public contracts and the PCO, we had to introduce measures to better describe our domain. For instance, the concept of tender was not fully expressed in the data model adopted in XML files, since there are only information about participants (inclusive of VAT numbers and company names), but not information related to offering services and prices. Nevertheless, the tender is one of the most important entity in the PCO to link the bidders to the public contract. For these reasons, during the conversion to linked data (Section 4), we decided to create

[18] Such data quality metrics are defined by the International Organization for Standardization: ISO/IEC 2501

[19] The Public Contracts Ontology is available on GitHub platform at: `https://github.com/opendatacz/public-contracts-ontology`

tender entities in our knowledge graph using as identifier the VAT number of the participant and the CIG of the contract.

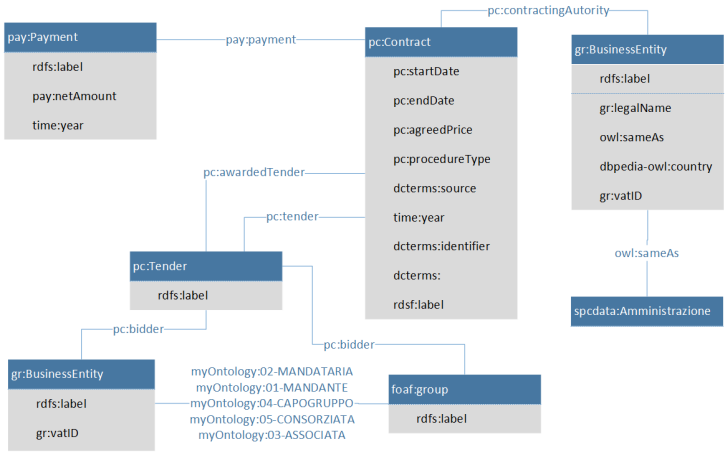

Fig. 5. A scheme of the data model used to build the ContrattiPubblici.org knowledge graph

4 Data Processing

When data derives from legacy databases, the publication of linked data is not always immediate [9]; data frequently comes from different sources and it needs to be gathered in a single file before proceeding with the conversion/translation into RDF triples (the so-called triplification) [3]. In the following section we show the process we used to obtain linked data as final result.

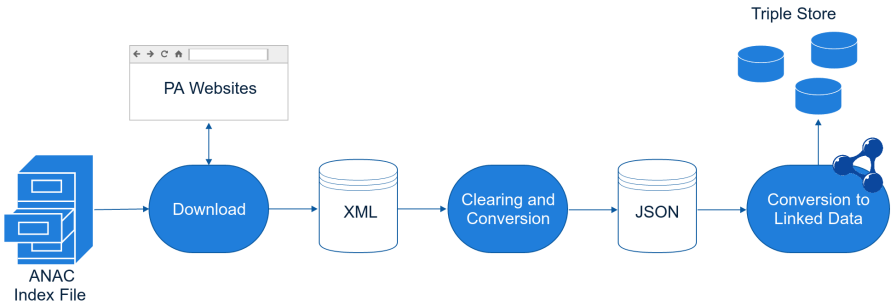

Fig. 6. Linked data conversion pipeline

4.1 Harvesting

As explained in Section 3, ANAC releases an index file that provides URLs of available XMLs, which are published on public administrations websites. Based on such index, the Download component tries to fetch data distinguishing between two different cases. In the first case, the component downloads and locally stores the XML containing public contracts data, with additional metadata related to the download outcome. In the second case, if fetched XMLs are indexes to files containing real data, the component is able to cross the links chain[20] and apply the download process shown in the first case. When the component is not able to recognize the expected schema of an XML with data or an XML index, it saves the file apart for a later manual check. In most cases, this means that either the URLs are wrong or the resource is not published according to an accepted format (e.g., it is a PDF file). In the worst cases, XML indexes are recursive, since they contain URLs that references to the XML index itself. For these reasons, we implemented some features in the Download component in order to manage this critical issue that threatens to undermine the entire pipeline. Moreover, during the download operation a lot of servers do not reply, for several reasons. We collected more than 10 different HTTP responses, which reveal how the quality of service over the 15000+ infrastructures of the Italian public administration might not be reliable.

4.2 Cleaning

By analysing the collected XMLs, we noticed that the quality of data is fragmented: different people with different systems led to inconsistencies and errors, as the example shown in Section 3.1. Therefore, we implemented a Cleaning and Conversion module that tries to guess potential errors for each field and attempts to correct data and define a standard format. In particular:

- dates are converted into the ISO 8601 format (YYYY-MM-DD);
- a digit check is performed on the CIGs (identifiers for each procurement) and on VAT numbers for detecting errors and verifying the syntactical correctness;
- agreed prices and payments, which are intended to be euro values, are casted to float number with two decimal digits;
- procedure types, which are fixed text categories, are checked with a function that calculates the similarity between strings. Such function tries to attribute unconventional values to one of those predefined categories.

Every value is analyzed and pushed in a result file serialized in JSON. If a value is modified by the Cleaning and Conversion component, both values, the original one and the guessed one, will be stored in this result file. Furthermore, a reference to the original XML file (the authoritative data) is included in such file.

[20] In some cases an index points to another index that finally might point to a file, or to another index

4.3 Conversion to Linked Data

After the Cleaning stage, we are able to convert the data into RDF using the N-Triples serialization[21]. During this conversion procedure, a component matches each field of the JSON file with the respective property (or relation) and data type from the Public Contract Ontology. When necessary, it also creates the needed entities (see the example of tender entities explained in section 3.2).

One issue we faced during the graph creation is the companies' labels management. While merging data from heterogeneous XML files, it is frequent to find different labels referring to the same company[22]. The creation of the labels triples is therefore handled by an algorithm that chooses the most common label referring to a company.

The last step in the conversion procedure is the so-called *interlinking*. Interlinking means declaring that an entity is *same as* another entity in another dataset, by adding new links to external resources and generating the so called knowledge graph. For this purpose, we chose the SPCData database[23], provided by the *Agenzia per l'Italia Digitale*, that contains the index of Italian public administrations. The knowledge graph is thus created by matching the VAT numbers of the original graph with the ones in the SPCData database.

After this procedure, the completed RDF file is pushed into a Triple Store that exposes data via a SPARQL endpoint.

5 Results

The semantic knowledge graph of ContrattiPubblici.org is published using the Virtuoso Triple Store[24], one of the most adopted technological solution to process and spread linked data. By using the SPARQL endpoint provided by Virtuoso, advanced users and robots are able to satisfy complex information needs. Table 1 details the total amount of RDF triples, entities, contracts, public bodies, companies, and external links to other datasets.

RDF triples	168,961,163
entities	22,436,784
contracts	5,783,968
public bodies	16,593
companies	652,121
links to external datasets	13,486

Table 1. Amount of public procurement information available on July 2017

[21] More information available at: `https://www.w3.org/TR/n-triples/`
[22] The differences may be minimal, as in presence of spelling errors, or even considerable
[23] More information available at: `http://spcdata.digitpa.gov.it/index.html`
[24] More information available at: `https://virtuoso.openlinksw.com/`

Due to data quality issues illustrated in Section 3.1, the number of companies presented in Table 1 is slightly overestimated. As explained in the next session, some future work will be dedicated to implementing further checks and expedients to merge different instances of the same company in a well-defined entity within the knowledge graph.

Despite these problems, in the context of transparency and open spending, it is possible to identify cases of public contracts with anomalies that require more investigation. With the following SPARQL query[25], for instance, advanced users and robots are able to get a list of 100 contracts in which the payment by the public body is more than doubled of the agreed price.

```
PREFIX pc: <http://purl.org/procurement/public-contracts#>
PREFIX payment: <http://reference.data.gov.uk/def/payment#>
SELECT ?contract ?amount ?agreedPrice WHERE {
    ?contract pc:agreedPrice ?agreedPrice.
    ?contract payment:payment ?payment.
    ?payment payment:netAmount ?amount.
    FILTER (?amount > 2*?agreedPrice)
} LIMIT 100
```

In addition to data analysis via SPARQL queries, users can exploit features of a human-consumption interface. For these reasons, we have developed a Web application that is available at: `http://public-contracts.nexacenter.org/`. Through a dedicated search form, users can enter the VAT number of a public administration, obtaining a visualization that shows different information. For example, a ranking of the top 10 beneficiaries on the basis of the total-allocated amounts (Figure 7), or a ranking on the basis of the total number of contracts (Figure 8) by means of histograms. Or even a view on contracts by means of a bubble diagram, allowing to compare very clearly the size of contracts put out to tender by the body (Figure 9). Through this kind of visualizations, an interested company could obtain an overview of tenders and contracts size, acquiring an increased knowledge on how to allocate its investments. Other features on the Web interface, including the search for individual contracts on the basis of keywords and new types of visualization, will be developed in the future.

6 Conclusions and Future Work

This contribution outlines opportunities in building a semantic knowledge graph based on linked data principles in the context of the Italian public procurement data. Despite the difficulties to tackle issues related to coverage and quality of open data sources, handling public contracts information in linked data enables novel avenues to evaluate the transparency of public administration and to create new business opportunities.

[25] The endpoint to perform the query is available at: `https://contrattipubblici.org/sparql`

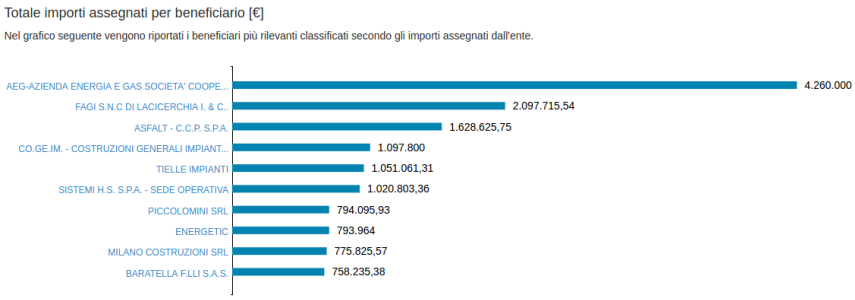

Fig. 7. Total amount of money (in Euros) assigned to a single beneficiary

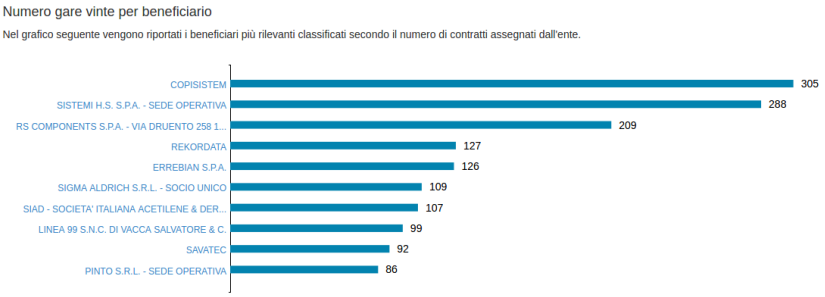

Fig. 8. Number of call for tenders won by a single beneficiary

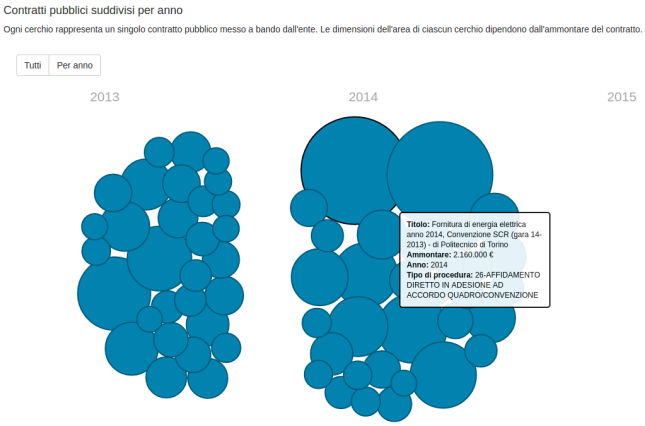

Fig. 9. Extent (in Euros) of the public call for tenders divided by year

As shown in the Results section 3.1, ContrattiPubblici.org semantic graph fosters the extraction of useful knowledge in different ways, ranging from SPARQL queries for complex analyses on data to rich and interactive visualizations. Furthermore, linked data principles lead to an enhanced interoperability across various data formats and Web applications, unleashing the full value of PSI.

Future work on ContrattiPubblici.org will most likely concentrate on the interlinking of the knowledge graph with more external datasets. To support this task, on the one hand, we plan to develop a component to automatize the process of Entity Identification, in particular for business entities. On the other hand, we will include metadata from DCAT-AP_IT[26] ontology to increase the semantic expressiveness of contracts data. Moreover, PCO described in section 3.2 will be extended with more details respect to the period of time of the contract, considering, for example, any interruption of the works, and the provenance of data. Further improvements for the research project are related to the addition of new data quality tests on procurement information, involving legal experts to control and validate the data after the cleaning process, and the development of new human-consumption interfaces.

References

1. Álvarez, J.M., Labra, J.E., Calmeau, R., Marín, Á., Marín, J.L.: Query Expansion Methods and Performance Evaluation for Reusing Linking Open Data of the European Public Procurement Notices, pp. 494–503. Springer Berlin Heidelberg, Berlin, Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-25274-7_50
2. Berners-Lee, T.: Putting government data online (2009)
3. Canova, L., Basso, S., Iemma, R., Morando, F.: Collaborative open data versioning: a pragmatic approach using linked data. In: CeDEM15 - Conference for E-Democracy and Open Governement. pp. 171–183. Edition Donau-Universität Krems, Krems (2015), http://porto.polito.it/2617308/
4. Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D.L., Hendler, J.A.: Twc data-gov corpus: incrementally generating linked government data from data.gov. In: WWW (2010)
5. Distinto, I., dAquin, M., Motta, E.: Loted2: An ontology of european public procurement notices. Semantic Web 7(3), 267–293 (2016)
6. Höffner, K., Martin, M., Lehmann, J.: LinkedSpending: OpenSpending becomes Linked Open Data. Semantic Web Journal (2015), http://www.semantic-web-journal.net/system/files/swj923.pdf
7. Martin, M., Stadler, C., Frischmuth, P., Lehmann, J.: Increasing the financial transparency of european commission project funding. Semantic Web Journal Special Call for Linked Dataset descriptions(2), 157–164 (2013), http://www.semantic-web-journal.net/system/files/swj435_0.pdf
8. Michail Vafolopoulos, M.M., et al., G.X.: Publicspending. gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives (jul 2012), http://www.w3.org/2012/06/pmod/pmod2012_submission_32.pdf

[26] More information on the Italian profile of the DCAT-AP defined in the context of ISA (Interoperability solutions for public administrations, businesses and citizens) program of the European Commission is available at: https://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0

9. Rowe, M., Ciravegna, F.: Data. dcs: Converting legacy data into linked data. LDOW 628 (2010)
10. Svátek, V., Mynarz, J., Wecel, K., Klímek, J., Knap, T., Nečaský, M.: Linked Open Data for Public Procurement, pp. 196–213. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-09846-3_10
11. Valle, F., dAquin, M., Di Noia, T., Motta, E.: Loted: Exploiting linked data in analyzing european procurement notices. In: Proceedings of the 1st Workshop on Knowledge Injection into and Extraction from Linked Data - KIELD 2010 (2010), http://sisinflab.poliba.it/sisinflab/publications/2010/VDDM10
12. Vetró, A., Canova, L., Torchiano, M., Minotas, C.O., Iemma, R., Morando, F.: Open data quality measurement framework: Definition and application to open government data. Government Information Quarterly 33(2), 325 – 337 (2016), http://www.sciencedirect.com/science/article/pii/S0740624X16300132