Removing Barriers to Transparency: A Case Study on the Use of Semantic Technologies to Tackle Procurement Data Inconsistency

Giuseppe Futia[1(✉)], Alessio Melandri[2], Antonio Vetrò[1], Federico Morando[2], and Juan Carlos De Martin[1]

[1] DAUIN, Nexa Center for Internet and Society, Politecnico di Torino, Turin, Italy
{giuseppe.futia,antonio.vetro,demartin}@polito.it
[2] Synapta Srl, Turin, Italy
{alessio.melandri,federico.morando}@synapta.it
https://nexa.polito.it
https://synapta.it/

Abstract. Public Procurement (PP) information, made available as Open Government Data (OGD), leads to tangible benefits to identify government spending for goods and services. Nevertheless, making data freely available is a necessary, but not sufficient condition for improving transparency. Fragmentation of OGD due to diverse processes adopted by different administrations and inconsistency within data affect opportunities to obtain valuable information. In this article, we propose a solution based on linked data to integrate existing datasets and to enhance information coherence. We present an application of such principles through a semantic layer built on Italian PP information available as OGD. As result, we overcame the fragmentation of datasources and increased the consistency of information, enabling new opportunities for analyzing data to fight corruption and for raising competition between companies in the market.

Keywords: Public procurement · Linked data · Data integration · Data consistency

1 Introduction and Motivations

Transparency refers to the principle according to which data related to functioning of government can be accessed and interpreted, without being predefined, preprocessed, and manipulated. As stated by Janssen [6], "adherence to this principle requires that the mechanisms for creating transparency are integrated in the heart of the government functions". Open Government Data (OGD) - i.e. data which is made freely available by public institutions to everyone - is a concrete step in this direction, playing a fundamental role in promoting transparency and accountability [9]. In addition, OGD improves the relationship among the

E. Blomqvist et al. (Eds.): ESWC 2017, Part I, LNCS 10249, pp. 623–637, 2017.
DOI: 10.1007/978-3-319-58068-5_38

government and citizens [11], who are enabled to be much more directly informed and involved in data driven decision-making[1].

Open Data on Public Procurement (PP), namely the procurement of goods or services on behalf of a public authority, is a specific area of the OGD characterized by big potential for increased openness of government information and incentives for supporting business activities. As reported by the Organisation for Economic Co-operation and Development (OECD), around US$ 9.5 trillion of public money is spent each year by governments procuring goods and services for citizens[2]. Furthermore, PP transparency is a crucial toolset to identify problems that arise from corruption, promoting competition and growth: according to the Transparency International Slovakia initiative[3] "reforms in procurement that included contract publication led to an increase in bids from an average of 2.3 per public tender in 2009 to 3.6 in 2013". In other words, as argued by Svátek [10], PP information is able to unify *public* needs and *commercial* offers: it enables a lively context to increase the interoperability between data models[4], methodologies, and sources independently designed within the two sectors.

Despite the tangible benefits that come with the publication of PP information as OGD, making data available does not automatically produce transparency. As underlined by Janssen [6], providing data alone is not sufficient: deep insights into the working of mechanisms to ensure that information can be easily accessible, processed, and interpreted are necessary to create transparency. Such reflections emerge from a conceptual framework called Big and Open Linked Data (BOLD), proposed by Janssen himself, that identify categories, dimensions, and sub-dimensions that influence transparency [6].

An obstacle to a comprehensive implementation of transparency through OGD is related to the fragmentation of existent open government datasets, in particular in the domain of procurement data[5]. As proposed by Tim Berners-Lee [2], linked data principles can be a *modular* and *scalable* solution to overcome the fragmentation in government data, increasing citizen awareness of government functions and enabling administrations to work more efficiently.

[1] The so-called "Participatory Governance" is one of the key aspect of the OGD mentioned by the Open Knowledge Foundation (OKFN). More information available at: http://opengovernmentdata.org/.

[2] More details available in the OECD blog post "Transparency in public procurement, moving away from the abstract": http://oecdinsights.org/2015/03/27/transparency-in-public-procurement-moving-away-from-the-abstract/.

[3] For more information, see: http://www.transparency.sk/.

[4] The ISA initiative of the European Commission represents a landmark for understanding different levels of data interoperability. More information available at http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

[5] Only 1/5 of total public expenditure on goods and services is published with rules complying with the EU Directives, for an estimated value of 420 billion. It means that "the bulk of total public expenditure on goods, services, and works is not organised in accordance with EU procurement legislation". See http://ec.europa.eu/internal_market/publicprocurement/docs/modernising_rules/executive-summary_en.pdf.

The data integration process, in fact, allows to identify cases of discrepancies respect to the knowledge base [13] of Italian procurement data. Some of these inconsistencies, in fact, can be only detected and fixed exploiting the crucial element of linked data principles, namely the interlinks between different data sources: accessing and retrieving data from an authoritative and reliable source, it is possible to reduce contradictions in the legacy dataset and enable opportunities to get less misleading results when the database is queried.

This article presents a solution based on linked data and semantic technologies to overcome the fragmentation of existing datasets and to increase the consistency of information: we process and transform PP data published on different websites of Italian administrations - in compliance with the Italian anti-corruption Act (law n. 190/2012) - relying upon linked data principles. The objective is to demonstrate how this approach has a fundamental impact to increase transparency of public bodies in the context of procurement data.

The structure of this contribution is the following. Section 2 describes the Italian context and gives an overall view of public procurement data made available by administrations. Section 3 explains current problems in terms of data quality of such data. Section 4 illustrates our approach for processing, transforming, and publishing procurement information as linked data. Section 5 reports results of the analysis on data quality issues. Section 6 presents a discussion on obtained results and limitations. Section 7 describes related work in the field of public procurement and spending information published according to linked data principles. The last section summarizes the work and present future advancements of the research project.

2 Study Context

In this Section we present the Italian legislative context according to which procurement data is published by public bodies and we describe the key characteristics of such data.

2.1 The Italian Legislative Context

The Italian Legislative Decree n. 33/2013 (DL33/2013) of March 14th, 2013[6] re-ordered obligations of disclosure, transparency, and dissemination of information by public administrations. According to specific requirements defined by the decree (Article 9 - DL33/2013), each body is required to create a specific section on its website called "Amministrazione Trasparente" (Transparent Administration). In this section, administrations provide details related to public procurement, with particular emphasis on procedures for the award and execution of

[6] See: http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2013-03-14;33!vig=. Notice that DL33/2013 has been recently amended by D.Lgs. 25 maggio 2016, n. 97 (DL97/2916), with a general tendency toward a more centralized publication of data - see, in particular, Article 9-bis - but no immediate impact on the publication requirements discussed in the paper at hand. See below footnote 7 for additional comments. Last visit on Nov. 2016.

public works, services, and supplies (Article 37 - DL33/2013[7]). Such data is published on the basis of a precise XML Schema Definition[8] (XSD) provided by *ANAC - Autorità Nazionale Anticorruzione* (Italian National Anti-Corruption Authority)[9], which has supervisory duties. After the publication on their websites, administrations transmit the link of the dataset to ANAC via certified mail. ANAC, at this point, performs a preliminary check and releases an index file (in JSON format), containing details related to the availability of data[10].

2.2 Source Data

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). 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[11]. The main structure of the XML file includes a section with the dataset metadata and a section containing multiple contracts, each of whom can be identified by the XML tag "lotto". The metadata section lists some information, including the first publication date and the last update of the dataset, the business name[12] of the contracting authority that spreads the dataset, the url of the dataset, and the license. The section containing data on contracts includes the following information: the identification code of the tender notice or CIG (that stands for Codice Identificativo Gara), the description of the tender, the procedure type for the selection of the beneficiary, the identification code and the business name of bidders (tender participants), the identification code and the business name of the beneficiary, the awarded amount, the paid amount, the dates of commencement and completion of works.

The second type of XML, instead, is an index that collects links to other XML files containing actual public procurement data[13].

3 Data Quality Problems

As described in Sect. 2.2, public contracts information is generated and spread on Italian public bodies websites. Due to diverse processes and tools adopted

[7] Following the aforementioned amendments by DL97/2016, a National Public Contracts Data Base -the BDNCP- is forthcoming, as described in the new Annex B of DL33/2013, but its creation will face all the problems described in the paper at hand - and possibly benefit from the approach that we suggest.

[8] XSD is a W3C recommendation that specifies how to describe an XML document.

[9] See: http://www.anticorruzione.it/.

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

[11] A representation of the XSD schema is available at http://dati.anticorruzione.it/schema/datasetAppaltiL190.xsd.

[12] A pseudonym used by companies to perform their business under a name that may differs from their legal name.

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

by administrations, the quality of PP data is extremely variable depending on the single case and it is impaired in terms of *accuracy*, *completeness*, and *consistency*[14].

Accuracy is defined as *the degree to which a data value conforms to its actual or specified value.* We distinguish between syntactical accuracy and semantic accuracy, which are defined in the following way:

– *Syntactical accuracy is defined as the closeness of the data values to a set of values defined in a domain considered syntactically correct.*
– *Semantic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered semantically correct.*

The definition of completeness is dependent on the perspective used:

– Computer system's point of view: *completeness is the extent to which all necessary values have been assigned and stored in the computer system.*
– End-user point of view: *completeness is the extent by which the data consumer's need is met.*

Consistency *refers to the absence of apparent contradictions within data.* Inconsistency can be verified on the same or different entities. In the context of XML data that refers to a schema, integrity constraints are properties that must be satisfied by all instances of a database schema. Although integrity constraints are typically defined on schemas, they can at the same time be checked on a specific instance of the schema that presently represents the extension of the database.

3.1 Interdependence Between Quality Metrics

Although there are different metrics to assess the quality of data as shown in the previous Section, such metrics are closely interdependent. In the procurement domain, for instance, a contract could present issues like bad comma position in a payment value (accuracy) or the lack of the payment field (completeness). Both errors have a direct impact on the consistency of information: contradictions inevitably occur when we analyze the total amount of expenditure resulting by several XML files that report data of an ongoing contract.

For these reasons, although the focus of the article concerns the consistency of the information, Sect. 5.2 proposes also a comprehensive analysis of the data quality in terms of accuracy and completeness.

3.2 Focus on Data Consistency

Certain types of consistency problems directly emerge analyzing contracts data collected in a single XML file. We report here 3 examples of such inconsistencies:

[14] Such data quality metrics are defined by the International Organization for Standardization: ISO/IEC 25012.

- contracts in which the beneficiary is more than one;
- contracts in which the amount of money is paid, but no recipient is present in the data;
- contracts in which the sum reported as paid is greater than the sum initially awarded to the beneficiary.

Other types of inconsistencies manifest themselves only after merging data contained in different sources. The following cases are real examples of inconsistencies with Italian PP data:

- business entities with more than one business name;
- CIGs that identify more than one contract;
- incoherent payments among different versions of an ongoing contract.

The aforementioned issues represent a significant barrier to achieve transparency, because the results obtained by querying the dataset are likely to be inconsistent and misleading. For example, consider a citizen trying to access the effective business name of a contracting authority identified by the id “00518460019” (this is VAT number of the “Politecnico di Torino”). If the data quality in terms of semantic accuracy is poor, such id could be associated with wrong business names like “Politecnico di Milano” and/or “Politecnico di Bari”. So, when the citizen performs a search using the business name as search key, he obtains an inconsistent result. The same problem happens when he wants to get details of a contract identified by a CIG: in this case, he is likely to get discordant values. Moreover, incoherent values of payments are not deductible from a single XML file, because errors emerge by analyzing the evolution of the contract data published in different years.

Section 5.2 show results of the analysis from the first kind of inconsistencies, that directly emerge analyzing contracts data collected in a single XML file.

In Sect. 6, instead, we demonstrate why reducing fragmentation of data with semantic technologies is an essential element to tackle the second type of inconsistencies, and we explain how some of these errors can be fixed exploiting linked data.

4 Applying the Linked Data Approach

In this Section we illustrate all stages of the approach to publish Italian PP according to linked data principles. Each stage is accomplished by means of different software components and resources that are shown in Fig. 1.

4.1 Harvesting of XML Files

The task of data harvesting is assigned to the Downloader component (3), that exploits the index file provided by ANAC[15] (1). On the basis of URLs contained in this file, the data fetching process from public bodies' websites (2) is able

[15] The index file provided by ANAC is available at http://dati.anticorruzione.it.

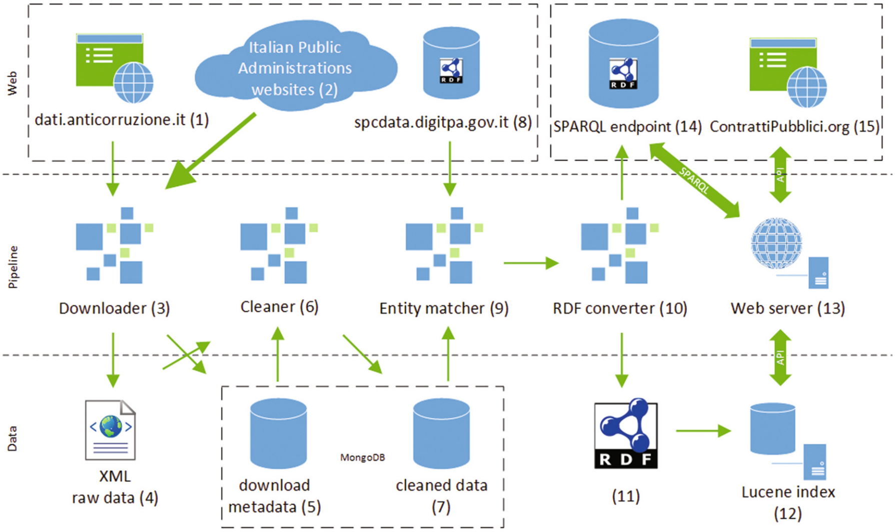

Fig. 1. Architecture for processing and publishing Italian PP as linked data

to manage two different cases. In the first case, the component gets XML files containing PP data and store them locally (4). Download metadata and the local path of each XML are stored in a MongoDB NoSQL database (5), in order to facilitate accessibility to such information and keep track of any duplicate file. In the second case, in which the XML contains links to other XML files, the component is able to cross the links chain[16] and performs the download process shown in the first case. When the component is not able to recognize the expected XML schema (as actual procurement data or as index) the file is stored in a dedicated directory of the file system for a later manual check. This problem occurs when 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 reference to the XML index itself. For these reasons, we implement some features 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: we collected more than 10 different HTTP responses, which reveal how the quality of service over the 15,000 infrastructures of the Italian public administration might not be reliable. Results of the download process are reported in Sect. 5.1.

4.2 Cleaning of Procurement Data

The next step to the data harvesting is performed by the Cleaner (6). During this stage procurement information is extracted from XML files and each contract

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

is processed and stored as unique document in an instance of the MongoDB database (7). Analyzing such data we evaluate the magnitude of data quality in terms of accuracy, completeness, and consistency (results of such evaluation are available in Sect. 5.2).

When we detect new errors we progressively improve the Cleaner component. Every time we apply a specific fix on data regarding to a contract, we add some metadata to the related MongoDB document: we preserve the original data and we compile a specific field called "errors" in which we describe the identified issue. For instance, if we encounter a bad format for the date value (dd-mm-yyyy), we transform such value in the correct format according to ISO 8601 (in our case yyyy-mm-dd), preserving the original data and saving in the "errors" field the following string: "bad date format" (Sect. 5.2 reports adopted solutions for the most common procurement data quality issues).

4.3 Public Contracts Ontology

In order to publish Italian public procurement according to linked data principles, we use the Public Contracts Ontology (PCO) developed in the context of the Czech OpenData.cz initiative[17]. According to its authors, this ontology describes "information which is available in existing systems on the Web" and "which will be usable for matching public contracts with potential suppliers" [4]. Therefore, the goal of the PCO is to offer a generic model for describing public contracts, without providing details of the public procurement domain, that are specific to fields and countries.

In the PCO domain, a call for tenders is submitted for the award of a public procurement contract. Therefore, we map XML fields and data described in Sect. 2.2 into entities, classes, and relations provided by the PCO. Figure 2 shows the data model adopted for publishing Italian PP as linked data. Although there is a significant degree of overlap between the XSD that describes the schema of source data and the PCO, we have to introduce additional elements 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, but not details related to offering services and prices. Nevertheless, the tender is one of the most important entity in the PCO to link participants to the public contract. For these reasons, during the conversion to linked data (Sect. 4.4), we create tender entities using as identifier the identification code of the participant and the CIG of the contract.

4.4 Triplification and Interlinking

After the cleaning stage, we convert contracts data stored in the MongoDB instance into RDF using the N-Triples serialization[18]. The component that

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

[18] More information available at: https://www.w3.org/TR/n-triples/.

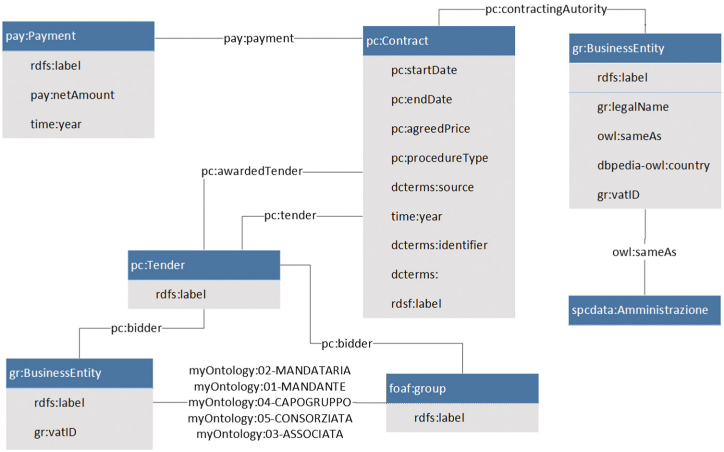

Fig. 2. Schema that describes Italian PP in the linked data domain

performs this task is called RDF converter (10), that maps fields and data values of the contracts into properties, entities, and literals defined by the PCO.

Before completing the triplification process, the Entity matcher component (9) performs the so-called *interlinking* stage. For our application, to improve the consistency of the information, we interlink public bodies listed in Italian PP to public bodies gathered from the SPCData database[19], provided by the *Agenzia per l'Italia Digitale* (8), that contains the index of Italian public administrations. The matching between entities are created using the identification code contained in both datasets. After the interlinking step, the final RDF file (11) is pushed into a Triple Store that exposes data via a SPARQL endpoint (14).

As shown in Fig. 1, the RDF file is also published in a Lucene[20] index (12) to enable full-text search features and data published within the endpoint can be queried by a Web server (13) to populate a Web interface[21] (15).

5 Results

In this Section we report results obtained with stages described previously in order to reduce data fragmentation and to improve data quality, in particular the consistency of information.

[19] More information available at: http://spcdata.digitpa.gov.it/index.html.

[20] Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. More information available at: http://lucene.apache.org/core/.

[21] The ContrattiPubblici.org project developed by Synapta Srl aims at demonstrating the opportunities for transparency and business of public procurement spread according to linked data.

5.1 Harvesting Results

With the approach described in Sect. 4, we integrate information coming from more than 300,000 XML files published by 15,000 public bodies. Table 1 reports details of the harvesting phase (see Sect. 4.1) that was accomplished in 4 different periods, starting from May 2015[22]. The download process has been carried out at different times for two reasons. The first reason is that ANAC releases the index file containing references to XML files in February of each year. The second reason is related to problems about servers uptime, which inevitably impacts on the availability of XML files. In order to tackle the last problem, on each harvesting cycle we try to download all XMLs, generating the duplication of files, that we manage during the data cleaning process.

Table 1. Number of downloaded XML files of procurement data in different periods of time

	May 2015	Nov 2015	Feb 2016	Nov 2016
URL requested	-	-	207.674	271.664
Downloaded files	205.415	184.738	201.451	252.246
Valid XMLs	199.341	180.609	197.338	247.881

5.2 Quality Problems Addressed

During the cleaning phase of PP (see Sect. 4.2), each contract is stored as single document within a MongoDB instance, enabling a first level of analysis on quality issues in terms of accuracy, completeness, and consistency of data. Table 2 shows, for each field of the contract, the type of data quality issue, the occurrence of such issue (in percent), the adopted solution (where available). The 41,65% of all contracts (almost 6 million in total) presents at least one of these issues. Analyzing in particular data inconsistency issues mentioned in Sect. 3.2, we discover that contracts in which the beneficiary is more than one correspond to 1.78% of all contracts; contracts in which the amount of money is paid, but no recipient is present in the data correspond to 4.30% of all contracts; contracts in which the sum reported as paid is greater than the sum initially awarded to the beneficiary correspond 5.96% of all contracts.

Exploiting the linked data principles, we build a semantic layer on procurement data to reduce fragmentation and to identify further inconsistencies. The dimension of the dataset built according to linked data principles is available in Table 3. Such dataset is published using the *Virtuoso Triple Store*[23] and can be queried via SPARQL endpoint[24].

[22] Some information is missing because in 2015 we did not store the number of requested URLs.

[23] More information available at: https://virtuoso.openlinksw.com/.

[24] The SPARQL endpoint on public procurement data is available at: https://contrattipubblici.org/sparql.

Table 2. Accuracy, completeness, and consistency degree in PP data

Field	Error	Occ. (%)	Solution
Completeness			
Start date	Missing	12.25	Nothing
End date	Missing	21.61	Nothing
Agreed price	Missing	0.06	Nothing
Payment	Missing	0.20	Nothing
Procedure type	Missing	0.11	Nothing
Business entity ID	Missing	1,05	Hash value
Accuracy			
Identifier	Syntactic errors	0.96	String cleaned
	Semantic errors	5.83	Hash value
Start date	Semantic errors	1.36	Nothing
End date	Semantic errors	2.00	Nothing
Agreed price	Syntactic errors	0.94	String cleaned
	Semantic errors	0.23	Nothing
Payment	Syntactic errors	0.76	String cleaned
	Semantic errors	0.65	Nothing
Procedure type	Syntactic errors	2.81	Optimal string match
Business entity ID	Semantic errors	1,08	Hash value
Consistency			
Start date	Non standard format	5.63	Uniformed to ISO 8601
End date	Non standard format	5.20	Uniformed to ISO 8601
Beneficiary	More than one beneficiary	1.78	Nothing
Payment	Payment without winner	4.30	Nothing
	Greater than awarded price	5.96	Nothing

Table 3. Characteristics of Italian procurement information published as linked data

Dimension	Value
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

6 Discussion on Inconsistency Issues

As explained in Sect. 3.2, there are some inconsistencies that are visible only after completing a data integration process. We resume 3 different reported cases:

- business entities with more than one business name;
- CIGs that identify more than one contract;
- incoherent payments among different versions of an ongoing contract.

The first of this case can be detected with the following SPARQL query:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gr: <http://purl.org/goodrelations/v1#>

SELECT (COUNT(DISTINCT ?be)) WHERE {
  {
   SELECT DISTINCT(?be) WHERE {
    ?be rdfs:label ?label .
    ?be a gr:BusinessEntity .
   }
   GROUP BY ?be HAVING (count(*)>1)
  }
}
```

This issue can be fixed exploiting the most important feature of linked data, namely the interlink among different datasets. In fact, we obtain a unique business name on a subset of business entities (in our case the contracting authorities), building links to the Italian public administration index of SPCData, shown in Sect. 4.4. From this dataset, exposed as linked data, we can get the official business name of contracting authorities, using as primary key their identification code (in our domain the VAT number of the contracting authority). In this way, we improve the consistency of information for a subset of business entities and we enable the opportunity to obtain valuable results.

The second case consists in the duplication of CIG for different contracts and can be detected with the following SPARQL query:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX pc: <http://purl.org/procurement/public-contracts#>

SELECT (COUNT(DISTINCT ?contract)) WHERE {
  {
   SELECT DISTINCT(?contract) WHERE {
    ?contract dcterms:identifier ?CIG .
    ?contract a pc:Contract .
   }
   GROUP BY ?contract HAVING (count(*)>1)
  }
}
```

The solution to this issue is generating a hash value, avoiding ambiguity due to duplicate CIGs, to build contracts URIs. In this way, we separate different contracts, misidentified by the same CIG, in different entities: we build the URI through a hash value generated combining the identity code of the contracting authority, the awarded amount, and the procedure type mentioned in the contract. The user is therefore able to detect this kind of error and he can semantically distinguish different contracts identified by the same CIG. Nevertheless, we need more context information to establish which is the correct CIG attribution for a specific contract.

The last problem is tracking incoherent payments published in different XML files of ongoing contracts. This problem is currently not solvable with our approach, because we lack information on provenance and reliability of different sources. In Sect. 6.1 we deepen this kind of limitation.

6.1 Limitations

The limitation of our approach is related to the absence of context information (in particular the reliability of the provenance) published in a structured way as linked data. In fact, we resolve inconsistency cases related to multiple business names for contracting authorities, because we can entrust an authoritative source (the SPCdata repository) that eliminates wrong business names in terms of semantic accuracy. Conversely, in the case of duplicate CIGs for different contracts and incoherent payments among different versions of an ongoing contract, we are not able to detect which is the invalid data because we can not support our analysis with an authoritative source. In other words, we are not in possession of the necessary data to adopt informed decision to govern the processing of information and consequently improve the quality of data. These limitations negatively impact on the level of achievement of transparency, because the citizen access to incomplete and controversial information.

7 Related Work

In this Section, we report contributions of procurement and spending data transformed and published according to linked data principles. This domain has already been addressed by several research projects, however a comprehensive work on Italian procurement data is not addressed yet. Furthermore, at the best of our knowledge, an analysis on procurement data consistency exploiting semantic technologies to improve transparency has not yet been accomplished.

One of the most important contributions in this domain is the LOD2 project, since it systematically addresses many phases of procurement linked data processing [10]. There are several other notable initiatives: the TWC Data-Gov Corpus [3], Publicspending.gr [8], The Financial Transparency System (FTS) project [7], Linked Spending [5], LOTED [12] and MOLDEAS [1].

In particular, the TWC Data-Gov Corpus gathers linked government data on US financial transactions from the Data.gov project[25]. This project exploits

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

a semantic-based approach in order to incrementally generate and enhance data via crowdsourcing. Publicspending.gr has the objective of interconnecting and visualizing Greek public expenditure with linked data to promote clarity and enhance citizen awareness through easily-consumed visualization diagrams. The FTS project of the European Commission contains information about grants for EU projects starting from 2007 to 2011, and publishes such data as linked data. Exploring this 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 conversion to linked data of information published by the OpenSpending.org, an open platform that releases public finance information from governments around the world. The project uses the DataCube vocabulary[26] to model data in order to represent multidimensional statistical observations. LOTED[27] is focused on extracting data from single procurement acts and aggregating it over a SPARQL endpoint. Finally, MOLDEAS presents some methods to expand user queries to retrieve public procurement notices in the e-Procurement sector using linked data.

8 Conclusion and Future Works

In this article we present an approach to tackle fragmentation of Italian procurement data and to improve consistency of such data. Both these issues represent a significant barrier to achieve a full transparency, because user and robots that query the datasets risk to obtain partial, inconsistent, and misleading results. As shown by our use case, some inconsistency problems already emerge analyzing data in XML files released by administrations. Among these issues we mention cases in which a contract presents more than one beneficiary; contracts in which the amount of money is paid, but the beneficiary is not present; contracts in which the payment is greater than the initially-awarded amount.

Nevertheless, in order to reduce data fragmentation and detect inconsistency cases that only emerge by integrating information, linked data and semantic technologies play a fundamental role. As demonstrated in this article, our approach allow user to interact with a single access point to information (the SPARQL endpoint), to detect inconsistencies cases, giving the opportunity to resolve some of them. Among these issues we mention: business entities with more than one business name; CIGs that identify more than one contract; incoherent payments among different versions of an ongoing contract. Some of these issues can be fixed exploiting peculiar characteristics of linked data principles, according to which is possible to create interlinks between different datasets. Business names of contracting authorities can be fixed exploiting an authoritative data source for public administrations, such as the SPCData portal released by the Agency for Digital Italy. For what concerns duplicate CIG values, we build URI generating a hash value to distinguish contracts, avoiding problems of data loss or data overlapping. In the last case (incoherent payments in ongoing contract) we are

[26] DataCube vocabulary information: https://www.w3.org/TR/vocab-data-cube/.
[27] LOTED project website: http://www.loted.eu/.

not able to apply any fix because we do not have enough information to establish the correctness of the data according to the provenance.

Among the future works, we mention the development of automatic tools to detect and fix consistency problems among contracts published in different files. We want to explore possible ways to evaluate the provenance of data in order to improve the data processing stage and improve data consistency. Lastly, we want to develop advanced methods and dashboards in order to monitor the consistency degree of the data.

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. In: Lozano, J.A., Gámez, J.A., Moreno, J.A. (eds.) CAEPIA 2011. LNCS (LNAI), vol. 7023, pp. 494–503. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25274-7_50
2. Berners-Lee, T.: Putting government data online (2009)
3. Ding, L., DiFranzo, D., Graves, A., Michaelis, J.R., Li, X., McGuinness, D.L., Hendler, J.A.: Twc data-gov corpus: incrementally generating linked government data from data. gov. In: Proceedings of the 19th International Conference on World Wide Web, pp. 1383–1386. ACM (2010)
4. Distinto, I., d'Aquin, M., Motta, E.: Loted2: an ontology of european public procurement notices. Semant. Web **7**(3), 267–293 (2016)
5. Höffner, K., Martin, M., Lehmann, J.: Linkedspending: openspending becomes linked open data. Semant. Web **7**(1), 95–104 (2016)
6. Janssen, M., van den Hoven, J.: Big and Open Linked Data (BOLD) in government: a challenge to transparency and privacy? Gov. Inf. Q. **32**(4), 363–368 (2015)
7. Martin, M., Stadler, C., Frischmuth, P., Lehmann, J.: Increasing the financial transparency of european commission project funding. Semant. Web J. Spec. Call Linked Dataset Descr. **2**(2), 157–164 (2013)
8. Vafolopoulos, M., Meimaris, M., Alexiou, G., et al.: Publicspending. gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives. In: Using Open Data: Policy Modeling, Citizen Empowerment, Data Journalism. W3C, The European Commission (2012)
9. Stiglitz, J.E., Orszag, P.R., Orszag, J.M.: Role of government in a digital age (2000)
10. Svátek, V., Mynarz, J., Węcel, K., Klímek, J., Knap, T., Nečaský, M.: Linked open data for public procurement. In: Auer, S., Bryl, V., Tramp, S. (eds.) Linked Open Data. LNCS, vol. 8661, pp. 196–213. Springer, Cham (2014). doi:10.1007/978-3-319-09846-3_10
11. Ubaldi, B.: Open government data. OECD Working Papers on Public Governance (2013)
12. 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)
13. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semant. Web **7**(1), 63–93 (2016)