POLITECNICO DI TORINO
Repository ISTITUZIONALE

Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study

Original
Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study / Torchiano, Marco; Vetro', Antonio; Iuliano, Francesca. - STAMPA. - (2017), pp. 144-153. ((Intervento presentato al convegno IEEE 41st Annual Computer Software and Applications Conference (COMPSAC 2017) tenutosi a Torino nel July 4-8.

Availability:
This version is available at: 11583/2670400 since: 2017-09-29T09:29:31Z

Publisher:
IEEE Computer Society

Published
DOI:10.1109/COMPSAC.2017.192

Terms of use:
openAccess

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

16 June 2019

Preserving the benefits of Open Government Data by measuring and improving their quality: an empirical study

Marco Torchiano[x], Antonio Vetrò[y] Francesca Iuliano[z]

[x]Dept. of Control and Computer Eng. (DAUIN)
Politecnico di Torino
Torino, ITALY
marco.torchiano@polito.it

[y]Nexa Center for Internet & Society
DAUIN - Politecnico di Torino
Torino, ITALY
antonio.vetro@polito.it

[z]Target Reply
Milan, ITALY
francesca.iuliano@studenti.polito.it

***Abstract*—Context: Open Government Data (OGD) represent an invaluable resource for enabling active citizenship. A significant example is represented by the mandatory data that Italian public administrations (PAs) are required to publish concerning their contracts. Nevertheless, a low quality of data provided by PAs could hamper the prospect of citizen involvement.**

Goal: Our objective is to define a set of basic metrics for public contracts OGD on the basis of the ISO SQuaRE standards family, with the goal of enabling the automated evaluation of dataset quality.

Method: We started with the metrics defined in the ISO 25024 standard and adapted them to the data schema of the OGD under evaluation. We assessed the results by looking at the issues revealed by the metrics applied to the data released by a pool of PAs.

Results: We were able to define a set of metrics, and apply them to to the datasets released by 12 distinct organizations. The metrics allowed us to identify several quality issues that limit the reuse of OGD from the citizen perspective.

Conclusions: The metrics we develop are able to identify quality issues and are suitable to perform an initial automated assessment of OGD datasets. This exercise also support the generality of the ISO 25024 quality measures.

***Keywords*-Data Quality; Data Quality Measurement; Open Government Data;**

I. Introduction

The idea underlying behind the Open Data is that data should be freely available to everyone to be used and re-published as they wish and without the restrictions from copyright or other mechanism of control [1]. The objective of the Open Data movement is similar to that one of other "Open" movements such as Open Software, Open Content and Open Access: the "Open" feature is supposed to foster collaboration, creativity and innovation in society [1], by promoting information exchange, knowledge, freedom of thought and benefits for the community.

There are many kinds of Open Data that have different potential uses and applications, such as: cultural, science, finance, statistics, weather, environment, transport. In the Public Administration field, Open Data is often called "Open Government Data" (OGD): the public sector is one of the major producers and holders of information, which ranges, e.g., from maps to companies records [2].

OGD are published with the aim of improving the transparency and the participation of the citizens in the administration of their nation: as a matter of fact, in recent years, the amount and variety of open data released by public administrations across the world has been tangibly growing [2], accompanied by increased political awareness on the subject [3]. Releasing public sector information as Open Data can provide considerable added value to society at large, meeting a demand coming from all kinds of actors, ranging from companies to Non-Governmental Organizations, from developers to simple citizens. Many suggest that wider and easier circulation of public datasets could entail interesting (and even unexpected) forms of reuse, for scientific [3] or commercial purposes [4]. In addition, OGD can improve transparency of public institutions [5] [6]: they allow an improvement of relationship among the government and citizens, who are enabled to be much more directly informed and involved in data driven decision-making.

Considering this potential, legal and technical openness of datasets is not sufficient, by itself, to create a prolific reuse ecosystem [7]: failures in managing the complexity of such large amount of data [8] and providing good quality information might impair not only the reuse of the data, but also the usage of the institutional portals [9]. For instance, in many cases OGD are not accurate [10], difficult to integrate with other data source (e.g., see examples on data from the two US chambers in [11]), incomplete or erroneous [12]. Hence,

[1]The Open Knowledge Foundation refers to Open Data as data that "can be freely used, modified, and shared by anyone for any purpose", see http://opendefinition.org/ last visited on Nov 26, 2016.

[2]See http://index.okfn.org/ last visited on Nov 26, 2016.

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

quality problems in OGD may easily impair all its potential reuses and those spillovers to society, which we listed above.

The paper at hand addresses this problem. By adopting a software engineering perspective and building on our previous work [13], we conducted an empirical analysis on the quality of contracts data from Italian Public Administrations: given the importance of such data especially in a historical period where policy and media are focusing great attention to spending of public money, we have chosen to analyse their quality and to examine the capability of a measurement approach derived from the software quality field.

II. Background and related work

The attention to Open Data quality has risen over the recent years. One of the best-known works in this field belongs to Tim Berners-Lee, who proposed a deployment scheme entitled "5 stars open data" [14]. This deployment scheme consists of five incremental quality requirements that are represented as stars. While this scheme indeed expresses one of the aspects of data quality, it focuses only on this one aspect, the format used to publish the data; thus cannot by itself be used to assess the total quality of a dataset. In 2007, a more all-around set of principles was produced by a group of Open Data and Internet experts who gathered under the moniker "Open government working group". The original set of principles contains eight rules in total, which state that any Open Data must be: Complete, Primary (as collected at the source), Timely, Accessible, Machine processable, Non-Discriminatory (available without registration), Non-Proprietary (in terms of format) and License-free. The original list has since then been extended with seven more rules, stating that the data must be: Online and free, Permanent (at a stable Internet location indefinitely and in a stable data format for as long as possible), Trusted, Documented, Safe to open, Designed with public input and there must exist a Presumption of Openness [15]. These principles have laid the basis for the development of an assessment process for Open Data quality.

Several data quality models and methodologies have been presented in literature, which has been collected in a detailed way by Batini et al. [16]. In addition to the models collected by their survey, the Software Quality Requirements and Evaluation (SQuaRE) model [17] and Portal Data Quality Model (PDQM) [18] have been later developed.

A further model developed by Moraga et al. [19], titled SQuaRE-aligned Portal Data Quality Model (SPDQM), was later introduced, and had been selected as a reference for the empirical evaluations in our previous work [13], as it provided a wider set of data quality characteristics than the others (the SPDQM contains 42 characteristics -30 from PDQM, 7 from SQUARE, 5 characteristics were added after a systematic literature review-, organised in two viewpoints) and well adapted to our case study. Moreover, since the OGD analysed span heterogeneous domains, it was preferable to select the dimensions that addressed the intrinsic aspects of data quality. In this viewpoint, SPDQM contained the most complete set of characteristics (12) in comparison to the other models listed by Batini et al. [16].

Then, starting from seven intrinsic quality characteristics from SPDQM, in [13], we identified a list of 14 metrics to evaluate the quality of a sample of Italian OGD. In the same time of the publication of our work [13], the ISO released the 25024 standard [20], which contains a set of 63 metrics to measure the quality of data on the characteristics previously defined in the ISO 25012 [17]. There, the ISO25024 made the 25012 operable and, sharing the same approach developed in software quality, it gives substantial basis for a measurement based evaluation of the quality of data.

Regarding Open Government Data and their quality measurement, a few precedent attempts have been done, which we briefly summarise below.

Ubaldi [21] developed a large set of metrics at very heterogeneous points of view (e.g., political, organisational, technical), measuring the quality of data in terms of availability (e.g., as number of datasets and metadata available on a specific portal), demand (e.g., number of views per day), reuse (e.g., number of apps developed with the data). All metrics proposed, however, lay at portal level, and were not evaluated.

In [22], the authors analyzed 50 datasets from Italian OGD at various administrative levels (regions, provinces, municipalities) in terms of completeness, accuracy and timeliness. Although the measurements proposed are at dataset level, the evaluation is performed at portal level by aggregating the values computed on each dataset. The authors observed that about 40% of regions and municipalities portals were not complete, i.e. did not make available the data requested by law, against 26% of the provinces portals. Regarding accuracy, the percentage of documents opened in a not machine-readable format ranged between 40% and 55%. Similar percentages were reported for timeliness. While we believe that this work is very relevant to understand quality problems that affect OGD, it was not based upon a theoretical framework because dimensions were not uniquely defined: as a matter of fact, the computed completeness was actually defined as availability, while accuracy was related to the format of the documents instead of their content.

Atz [23] proposed the *tau metric* to capture the percentage of datasets up-to-date in a data catalogue, applying it to three different portals (World Bank, the UK data catalogue and the London data store). The author computed the metric on the retrieved datasets and then aggregated the obtained measurement to form a single indicator of Timeliness, which also discriminated between new release and minor updates. Results indicated that in two portals only about half of the datasets were updated according to their schedule and the nature of the contained data, while in the third one only one fourth did. Notwithstanding the different metrics construction, these findings are similar to those in [22].

Behkamal et al. [24] investigate Open Linked Data quality: the authors took as reference the ISO25012 standard data quality model and built a set of 20 metrics related to semantic and syntactic accuracy, uniqueness, completeness and consis-

tency. They verified the suitability of the proposed framework both with a theoretical validation and an empirical one. From the theoretical point of view, all of the metrics respect four out of five desirable properties, namely non-negativity, null value, symmetry and monotonicity, but not additivity. However, being additivity a special case of monotonicity, the authors state that the satisfaction of the monotonicity property makes them acceptable for their intended usage. The results of the empirical evaluation lead to the exclusion of four not discriminative metrics (ratio of syntactically incorrect triples, ratio of instances being members of disjoint classes, ratio of functional properties with different values, invalid usage of inverse-functional properties), and to the observation that a dataset with higher number of similar properties is highly likely to have more triples using these properties, while using similar properties have an inverse relation with the inconsistency of data values in a dataset.

These studies were relevant in demonstrating that measurement approaches can help improving the quality of OGD. Our previous and current work takes foundations from them and provides new evidence and details on quality issues affecting strategic and relevant Open Government Data (we describe them in III). Our contribution differentiate from previous ones in terms of granularity level (we assess the quality at single dataset level and not at portal level) and reference model (we explicitly refer to ISO 25012 and ISO 25024 theoretical framework). With respect to our work in [13], we address new technical problems due to the shift from tabular data to xml data, and we check the alignment to the ISO 25024, which was not yet published at that time.

III. Context and data collection

The Italian Legislative Decree n.33 of 14 march 2013[4] (DL33/2013) concerns the obligations of publicity, transparency and dissemination of information by public authorities. The decree makes explicit the function of public utility of Open information, indicated as "widespread forms of control on the pursuit of official duties and the use of public resources". With regard to publication requirements, the decree mandates the creation of a special section, called "Amministrazione Trasparente" (Transparent Administration) in the home page of the organization website where relevant information will be published. In the clause no.37 of DL33/2013, reference is made to the disclosure obligations related to public contracts and, in particular, it is expected that within January 31th of each year summary tables are published in an open standard format, allowing anyone to analyse and process the data for statistical purposes. In addition, administrations are required to transmit this information in digital format to the Italian Anti-Corruption Authority for the supervision of public contracts. The format chosen by the Authority, for the transmission of such data, is XML compliant to a pre-defined schema. Transmission is done by communicating the URL of publication of the data files.

[4]http://www.decretotrasparenza.it/wp-content/uploads/2013/04/D.Lgs_.-n.-332013.pdf last visited on Nov 26, 2016.

The full XML schema is reported in Figure 7 at the end of the paper, in summary it is structured as follows:

- A section with the metadata of the dataset
- A section containing a list of lots[5], more precisely:
  - a Each record corresponds to a lot. The record is tree-structured, that is, exist for each lot a number of child records that can report information with variable cardinality.
  - b Each lot consists of several elements.
  - c Each file can contain multiple lots.

After careful scouting of the data, we deemed the public contracts that Italian universities publish on their website of relevant interest. We extracted the data from the XML files provided by the Universities, and loaded them on a MySQL database (following a data model equivalent to the XML representation), to have higher flexibility in our computations. The code we used has been released as open source on the GitHub repository [6].

IV. Measurement and method

We evaluate the quality of the public contracts data on those quality characteristics defined in the ISO 25012 standard that were suitable to the type of data under study and that lended themselves to automatic computation. In the following we describe the quality dimensions selected as defined by the standard ISO 25012 and the relative motivations for deviations from the definition provided by ISO 25024 to fit our context. The metrics descriptions are reported in Table I.

a) Accuracy: 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 (Example: a low degree of syntactical accuracy is when the word Mary is stored as Marj).*
- *Semantic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered semantically correct. (Example: a low degree of semantic accuracy is when the name John is stored as George. Both names are syntactically accurate, because of the domain of reference in which they reside, but George is a different name.)*

For the specific case of public contracts, given the unavailability of an oracle for all correct values, it is impossible to establish the semantic accuracy of the data. Instead, we exploit the information about the domain of the data provided in the XML Schema to check whether values belong to the relative domain, that is we assess the syntactic accuracy of the data.

[5]A *lot* is a request for the procurement of specific products or services and the related acquisition

[6]https://github.com/xxxxx⟨blinded⟩xxxxxx last visited on Jan 20, 2017.

b) Completeness: 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. Completeness refers both to entity occurrences and to attributes of a single occurrence.*
- End-user point of view: *completeness is the extent to which data are sufficiently able to satisfy user's stated needs from quantitative point of view. Completeness includes also the capability of data to represent the context observed by users.*

As reported in [16], completeness in a relational model can be characterised with respect to:

- Presence or absence and meaning of the null values.
- The validity of one of the two assumptions called open world assumption and closed world assumption.

For the case of public contracts the model without null values and with open world assumption would provide a high accuracy in the evaluation of the completeness but cannot be used since we do not have a reference relation. For this reason we use the model without null values and closed world assumption and in particular we evaluate the tuples and attributes completeness, i.e.:

- Tuple completeness: the completeness of a tuple with respect to the values of all its fields.
- Attribute completeness: the number of null values of a specific attribute in a relation.

c) Consistency: 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. Therefore it is possible to distinguish two main categories of integrity constraints:

- Intrarelational Integrity Constraints: regard single attribute or multiple attributes of a relation.
- Interrelational Integrity Constraint: involve attributes of more than one relation.

The data for public contracts allow the definition of many constraints. Some of them are specified in the XML Schema, some other are logical constraints that we have defined on the data after a careful study of the domain. In the evaluation of the data we have defined both intrarelational constraints and interrelational constraints as reported in Table I.

In addition, following the recommendations of the ISO 25024, it is possible to compute a risk of inconsistency by looking at duplicated values: duplication occurs when a real-world entity is stored twice or more in a data source. Of course, if a primary key consistency check is performed when populating a relational table, a duplication problem does not occur if the primary key assignment has been made with a reliable procedure. Indeed, the duplication problem is more relevant for files or other data structures that do not allow the definition of key constraints.

The duplication of data related to public contracts is evaluated on the values that if duplicated would make the provided data unreliable (see details on Table I).

V. Results

We applied the metrics to 123,702 lots, belonging to the 12 universities which provided the data. Figure 1 reports the number of lots in each organization we analyzed. We observe a wide range with roughly one order of magnitude between the smallest and the largest. Such variability can be partly explained by the size of the universities, the left-hand side of the figure reports the enrolled students in each university for the academic year 2013/14.

Due to the large number of measurements and metrics, we report here only the results that revealed interesting aspects or problems, that will be discussed in the next section (the reader might consult the complete set of results online [7]).

A. Accuracy and Completeness

Concerning the *Accuracy* and *Completeness* metrics, we report the measures related to the *cig* of the lot, that represents a unique identifier for the tender, the contractor selection method for the lot, and the participant's unique fiscal ID.

The *cig* is a crucial information because it provides the unique id of each contract. The measures of accuracy (red triangles) and completeness (blue circles) for *cig* are reported in Figure 2; we can observe that the percentage of complete values is generally high for all the organizations. While the accuracy level is less consistent: although the element *cig* is present in (almost) all the lots, in some cases it is outside the valid domain. In practice it happens to be a number with a number of digits other than 10 or it is blank. The most relevant case is that of UniTo, which published data that in more than 30% of the cases exhibits invalid *cig* and thus are no more suitable to uniquely identify a contract.

The Contractor Selection indicates the type of procedure followed to select the contractor – e.g. direct selection, public tender, etc. – and it can potentially be used by the control authorities to identify patterns of illegal award of contracts. A good level of accuracy and completeness of this information is fundamental to increase transparency of public contracts. Figure 3 reports the measures, we notice one case (UniMi) with 100% completeness together with a 0% accuracy: the element in the original file is always present but it is empty.

The accuracy and completeness for the Participants' fiscal ID - Figure 4 - (and for all the attributes of the participants) is not computed for UniMi. In all the files analysed for UniMi there isn't any participant, that is, all the lots have a successful tenderer but no information about the participants. In general, both the accuracy and completeness of the fiscal Id is high. PoliMi has about the 4% of inaccurate values due to the fact

[7] https://github.com/xxx⟨blinded⟩xxxxx.pdf last visited on Jan 20, 2017.

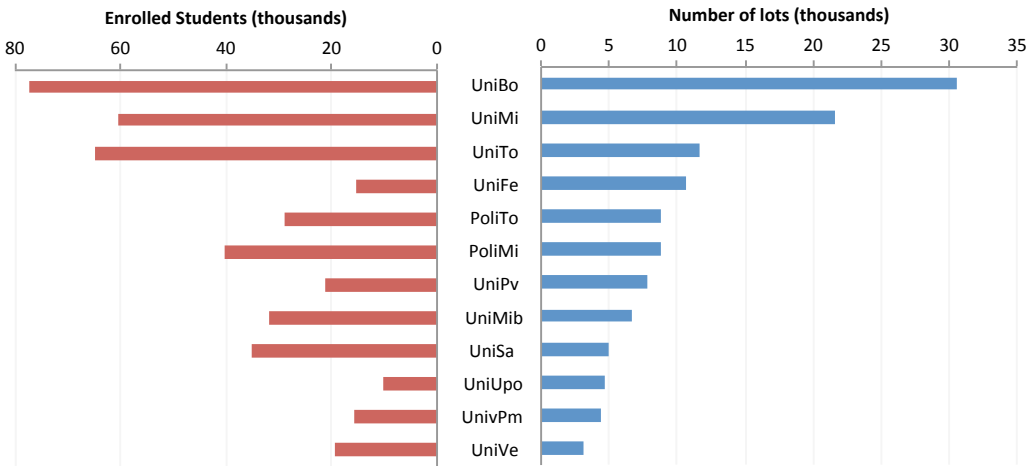

Fig. 1. Number of lots for each organization

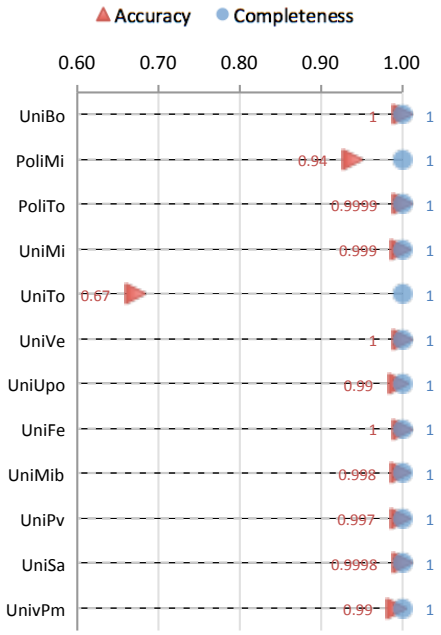

Fig. 2. Accuracy and completenes measures for CIG

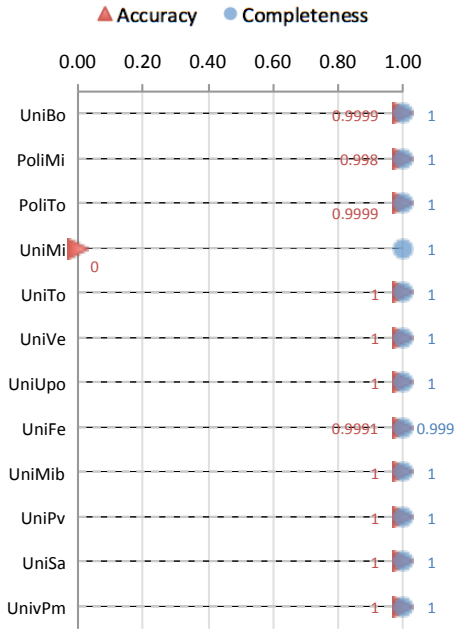

Fig. 3. Accuracy and completenes measures for Contractor Selection

that the values extracted from the XML files are either other than 11 or 16 digits or empty.

B. Consistency

Figure 5 reports the values for three inter-relational constraints indicators.

The leftmost one – *lot_has_participant* in Table I – shows the percentage of lots which have at least a participant. The most critical case is that of UniMi that has a percentage equal to zero, also UniTo and UniSa exhibit a very low consistency. This is due to the lack of information on the participants in the published files.

The middle indicator – *successfulTenderer_is_participant* in Table I – shows the percentage of lots for which the contractor is one of participants. The possible values are clearly limited by the previous indicator – if no participant is present then the contractor cannot be found among them –. We observe slightly smaller values indicating that even when the participants are listed, it may happen that the contractor is not one of them.

The rightmost indicator – *successfulTenderer_awardAmount* in Table I – shows the percentage of lots with a recorded payment larger than zero that reports a contractor. In the case of UniFe, more than 40% of the contracts report some money was payed but no recipient for the sum – the contractor – is present.

In addition, we report in Figure 6 one intra-relational metric – *isl_lt_et_ia* in Table I –; this indicator reports the percentage of lots in which the sum reported as paid is greater than the

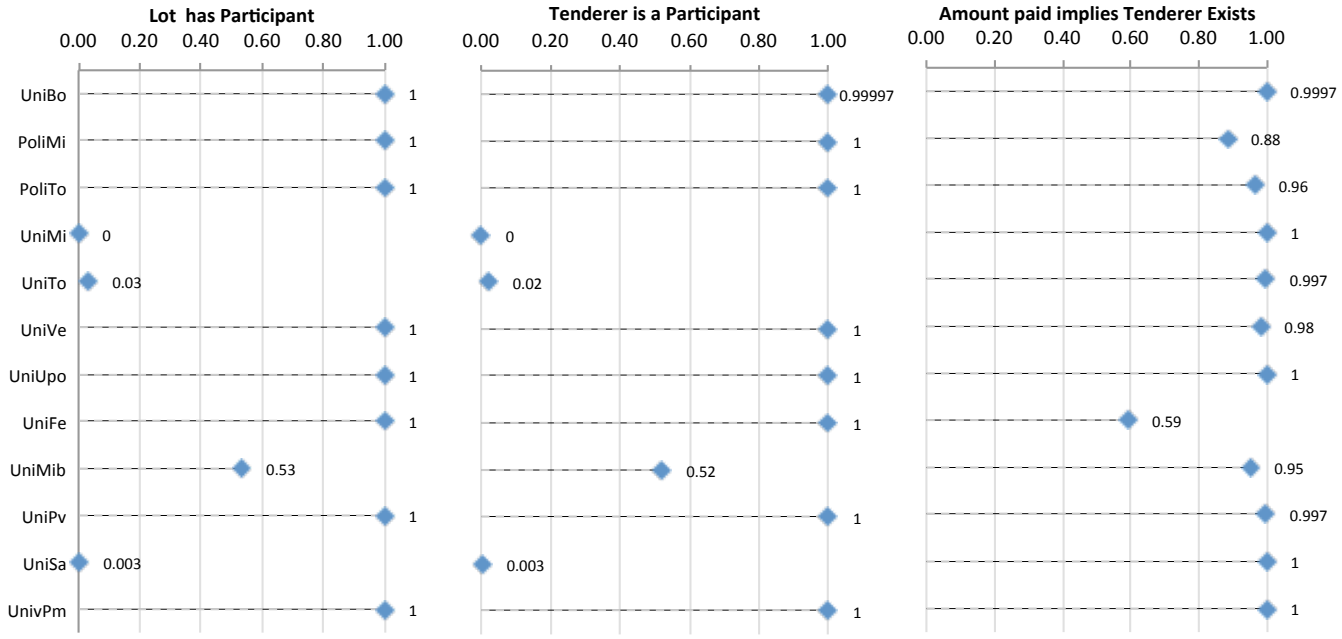

Fig. 5. Inter-relations constraints measures: *lot_has_participant*, *successfulTenderer_is_participant*, *successfulTenderer_awardAmount*

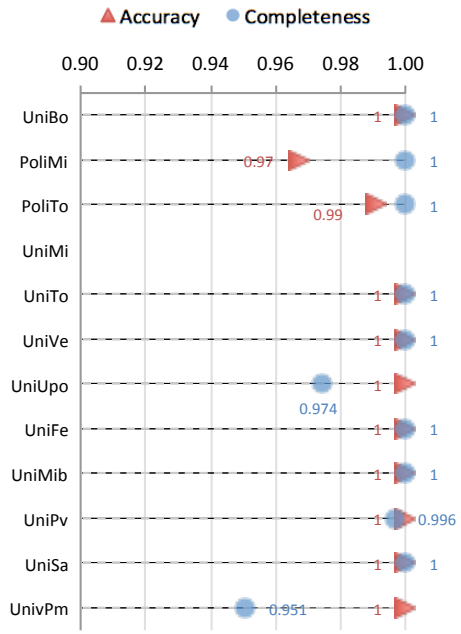

Fig. 4. Accuracy and completenes measures for Participant Fiscal Id

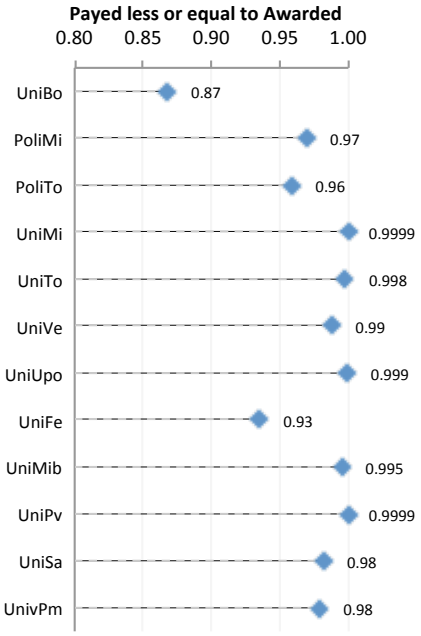

Fig. 6. Intra-relational constraint measure

sum initially awarded to the contractor. We observe that UniBo has around the 13% of the lots that violates the consistency constraint. For those lots a payment higher than the awarded amount was detected.

VI. Discussion

The results selected and presented in Section V have been focused on a few features that we deemed highly relevant for the reuse of the data on public contracts. The measures revealed a set of interesting problems: here we briefly discuss how the detected data quality problems affect the reuse of data and seriously endanger the transparency and spillovers that represent the ultimate goals of OGD.

Starting with the measurements on accuracy and completeness (reported on Figures 2, 3 and 4), it emerged that the data fields analyzed exhibit a generally high level along both

dimensions, though occasionally levels are not high.

For instance, regarding the *cig* – the unique identifier of contracts – we observed that the percentage of complete cells is rather high for all the organizations but, for some of them, the percentage of accurate cells is subject to wide variations. This means that, although the element *cig* is present in all the lots, in some cases its value lies outside the allowed domain. This happened in 33% of the cases for UniTo, which means that for that university about 4,000 lots out of 12,000 were invalid: any further reuse of these data will be misleading or incomplete in this specific case.

Concerning the *sceltaContraente* field, we noticed that universities payed much attention to providing this information: in fact, the percentage of complete and accurate cells for the *sceltaContraente* attribute is the 100% for almost all universities. However, we observed one case where the *sceltaContraente* element in the original file although always present is systematically left blank. Since this regards the procedure of contractor selection, in that case it would be impossible to carry out any analysis for identifying patterns of illegal awards of contracts; in fact this is an extremely relevant problem which affects public offices in Italy and represents one of the motivation that led to the legislation mandating the publication of the public contracts data.

Finally, for the *codiceFiscale* field, we found both cases of 100% accuracy and lower completeness and viceversa. Again, these measurements are the consequence of cases in which the field is present in the lots but left blank, or the values are other than the admissible ones. Being the *codiceFiscale* the Italian fiscal identifier for persons and organizations, any detailed analysis or representation of data would be impossible, and only aggregated results could be derived.

All the types of quality problems found with accuracy and completeness metrics could have been caused by a faulty manual data input, a defect in the software system regulating the information flow, or an incomplete integration of the different information systems managing the contract procedure. Whatever cause, however, an automatic measurement – such as the one we presented in this paper – could have detected the problems and potentially allowed avoiding their effects on reuse. In addition, simple instruments for data cleaning (e.g., Open Refine[8] , Data Cleaner [9]) can be useful in absence of domain specific tools.

Regarding the consistency dimensions, the checks on inter- and intra- relational constraints (Figures 5 and 6) revealed also relevant problems.

We observed that the percentage of lots with at least a participant is quite variable. The most interesting cases are UniMi, UniTo and UniSa with a percentage equal or close to zero. As a consequence, also the *successfulTenderer_is_participant* inter-relational constraint has been affected. These measures showed clearly that for all lots there was a contractor but no information on participants, which is a non-realistic situation. In practice, a citizen cannot trust any of the data on the contracts published by these three organisations: this corresponds to about 30 thousands lots, i.e. one third of all the data on contracts available from Italian universities.

Finally, we observed cases of expenses that overcame the awarded amount (e.g., 13% of lots at UniBo) or cases in which a grant winner is absent however the amount paid is different from zero (up to 40% at UniFe): again, this is not an admissible situation. Being UniBo the organisation with the highest number of lots, this portion of unusable data contributes to further increase the number of untreatable data estimated above.

[8]http://openrefine.org/ last visited on Nov 26, 2016.

[9]http://datacleaner.org/ last visited on Nov 26, 2016.

VII. Conclusions

We defined a set of metrics to empirically evaluate the quality of OGD on public contracts, on the basis of the quality measures defined in the ISO 25024 standard. The computation of the metrics starting from the XML files published by the PAs has been automated by means of an open-source piece of software. The metrics were then validated on the datasets published by the Italian Universities.

We were able to automatically analyse a large amount of data and to observe several problems affecting data accuracy, completeness and consistency. The problems were so relevant that just the subset of measurements presented in this paper invalids about a third of the whole Italian catalogue on universities public contracts. In fact, any reuse of such portion of data would lead to misleading or incomplete results, impairing the ultimate goals of Open Government Data: enabling active citizenship, fostering transparency in public administration and stimulate economical spillovers.

Are the quality problems reported in this work a result of illicit procedures ? Or, are they rather injected by faulty data integration/acquisition processes? The answer to this question does not concern neither the authors of this paper nor the scientific community at large. The point, however, is that in absence of a measurement approach on OGD, the civic and economical spillovers from reusing the data are *tout court* canceled. We provided empirical evidence in support on this statement, and produced two important contributions: (i) benchmark data on OGD quality and (ii) a measurement framework which is applicable to any other data on contracts published by Italian PAs, since the data format is standardised by the Authority for the supervision of public contracts.

Our next steps entail a renovated analysis on the updated data on public contracts and a more comprehensive benchmarking analysis on the whole landscape of Italian OGD.

Acknowledgment

We thank Lorenzo Canova, Federico Morando and Raimondo Iemma for their contributions during the work and in the dissemination phase.

References

[1] J. Hofmokl, "The internet commons: towards an eclectic theoretical framework," *International Journal of the Commons*, vol. 4, no. 1, pp. 226–250, 2010.

[2] G. Aichholzer and H. Burkert, *Public sector information in the digital age: between markets, public management and citizens' rights*. Edward Elgar Publishing, 2004.

[3] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Information Systems Management*, vol. 29, no. 4, pp. 258–268, 2012.

[4] G. Vickery, "Review of recent studies on psi re-use and related market developments," *Information Economics, Paris*, 2011.

[5] J. E. Stiglitz, P. R. Orszag, and J. M. Orszag, "Role of government in a digital age," *2000.*, 2000.

[6] B. Ubaldi, "Open government data," 2013.

[7] N. Helbig, M. Nakashima, and S. S. Dawes, "Understanding the value and limits of government information in policy informatics: a preliminary exploration," in *Proceedings of the 13th annual international conference on digital government research*. ACM, 2012, pp. 291–293.

[8] D. Natale, "Complexity and data quality," in *Poster e Atti Conferenza*, 2011, pp. 13–16.

[9] B. Detlor, M. E. Hupfer, U. Ruhi, and L. Zhao, "Information quality and community municipal portal use," *Government Information Quarterly*, vol. 30, no. 1, pp. 23–32, 2013.

[10] B. Allison, "My data can't tell you that," *Open Government: Collaboration, Transparency, and Participation in Practice. O'Reilly Media, Inc*, pp. 257–265, 2010.

[11] J. Tauberer, *Open government data*. Joshua Tauberer, 2012.

[12] A. Whitmore, "Using open government data to predict war: A case study of data and systems challenges," *Government Information Quarterly*, vol. 31, no. 4, pp. 622–630, 2014.

[13] B. A. List, "Xxxx blinded title xxxxxx," *Government Information Quarterly*, vol. X, no. X, 2016.

[14] T. Berners-Lee, "Linked data-design issues," W3C, Tech. Rep., 2006.

[15] VV. AA., "The annotated 8 principles of open government data. the 8 principles of open government data, 7 additional principles." Tech. Rep., 2014. [Online]. Available: http://opengovdata.org.

[16] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, July 2009.

[17] ISO/IEC, "25012 international standard: Systems and software engineering - software product quality requirements and evaluation (square)-data quality model," ISO/IEC, Tech. Rep., 2008.

[18] A. Caro, C. Calero, I. Caballero, and P. M., "A proposal for a set of attributes relevant for web portal data quality," *Software Quality Journal*, vol. 16, no. 4, pp. 513–542, 2008.

[19] C. Moraga, M. Moraga, C. Calero, and A. Caro, "Square-aligned data quality model for web portals," in *QSIC'09. 9th International Conference on Quality Software*, 2009, pp. 117–122.

[20] ISO/IEC, "Iso/iec 25024:2015 systems and software engineering – systems and software quality requirements and evaluation (square) – measurement of data quality," ISO/IEC, Tech. Rep., 2015.

[21] B. Ubaldi, "Open government data: Towards empirical analysis of open government data initiatives," OECD Publishing, Tech. Rep., 2013.

[22] A. Maurino, B. Spahiu, C. Batini, and G. Viscusi, "Compliance with open government data policies: an empirical evaluation of italian local public administrations," in *Twenty Second European Conference on Information Systems*, 2014.

[23] U. Atz, "The tau of data: A new metric to assess the timeliness of data in catalogues," in *Conference for E-Democracy and Open Governement*, 2014, p. 257.

[24] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," *Journal of theoretical and applied electronic commerce research*, vol. 9, no. 2, pp. 64–79, 2014.

TABLE I
CHARACTERISTICS AND METRICS USED IN THE EVALUATION

Dimension	**Acronym**	**Metric**	**Relation to ISO/IEC 25024**
Accuracy	pcvc	Percentage of cells with correct value (value belonging to the domain)	Same as Acc-I-1
Completeness	pcc	Percentage of complete cells	Same as Com-I-4
	pcrp	Percentage of complete tuples.	Adaption of Com-I-1
Consistency (Duplication)	dup_participant	Number of participant which are duplicated in each lot.	Adaptation of Con-I-3
	dup_tenderer	Number of successful tenderer which are duplicated in each lot.	Adaptation of Con-I-3
Consistency (Intrarelational-interrelational constraints)	cf_ife_partecipanti	Percentage of tuples that meet the following Intrarelational Constraint: codiceFiscale and identificativoFiscaleEstero in the partecipanti table must not be simultaneously not null.	Adaptation of Con-I-6
	cf_ife_aggiudicatari	Percentage of tuples that meet the following Intrarelational Constraint: codiceFiscale and identificativoFiscaleEstero in the aggiudicatari table must not be simultaneously not null.	Adaptation of Con-I-6
	cf_eq_zero_partecipanti	Percentage of cells that meet the following Intrarelational Constraint: codiceFiscale in partecipanti table must be different by zero.	Adaptation of Con-I-6
	cf_eq_zero_aggiudicatari	Percentage of cells that meet the following Intrarelational Constraint:codiceFiscale in aggiudicatari table must be different by zero.	Adaptation of Con-I-6
	check_on_date	Percentage of tuples that meet the following Intrarelational Constraint: dataInizio must be less recent than dataUltimazione in lotti table	Adaptation of Con-I-6
	isl_lt_et_ia	Percentage of tuples that meet the following Intrarelational Constraint: ImportoSommeLiquidate must be less than or equal to ImportoAggiudicazione in lotti table	Adaptation of Con-I-6
	lot_has_participant	Percentage of tuples that meet the following Interrelational Constraint: When a lot has a successful tenderer it must have at least one participant.	Adaptation of Con-I-6
	successfulTenderer_is_participant	Percentage of tuples that meet the following Interrelational Constraint: A successful tenderer of a lot must be a participant for that lot.	Adaptation of Con-I-6
	successfulTenderer_amountPaid	Percentage of tuples that meet the following Interrelational Constraint: When the successful tenderer is not present for a lot, the amount paid (importoSommeLiquidate) must be zero for that lot.	Adaptation of Con-I-6
	successfulTenderer_awardAmount	Percentage of tuples that meet the following Interrelational Constraint: When there is a successful tenderer the award amount (importoAggiudicazione) must be different by zero.	Adaptation of Con-I-6

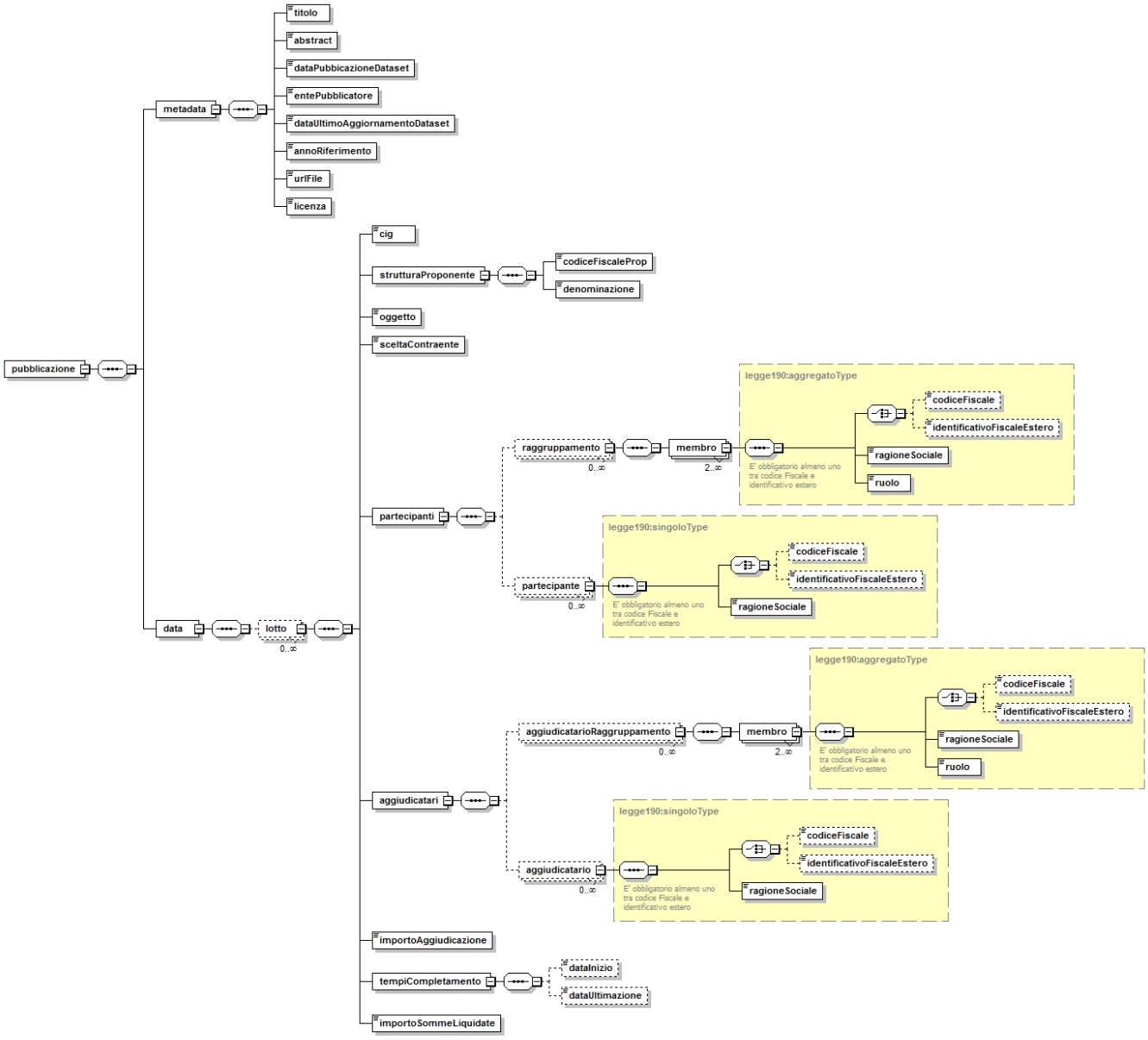

Fig. 7. XML schema for the publication of the public contracts data, taken from the technical specification available at http://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%20Assets/anacdocs/Servizi/ServiziOnline/AdempimentoLegge190/Specifiche%20Tecniche%20Legge%20190%20v1.2_finale.pdf last visited on Nov 26, 2016.