Open Data Quality Measurement Framework: Definition and Application to Open Government Data

Open Data Quality Measurement Framework: Definition and Application to Open Government Data

Antonio Vetrò

Nexa Center for Internet & Society,

DAUIN, Politecnico di Torino (Italy)

antonio.vetro@polito.it

Nexa Center for Internet & Society,

Politecnico di Torino

Via Pier Carlo Boggio, 65

10138 Torino

Torino

+39 011 090 5954

Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, Federico Morando

Politecnico di Torino (Italy)

name.surname@polito.it

ABSTRACT

The diffusion of Open Government Data (OGD) in recent years kept a very fast pace. However, evidence from practitioners shows that disclosing data without proper quality control may jeopardize datasets reuse and negatively affect civic participation. Current approaches to the problem in literature lack of a comprehensive theoretical framework. Moreover, most of the evaluations concentrate on open data platforms, rather than on datasets.

In this work, we address these two limitations and set up a framework of indicators to measure the quality of Open Government Data on a series of data quality dimensions at most granular level of measurement. We validated the evaluation framework by applying it to compare two cases of Italian OGD datasets: an internationally recognized good example of OGD, with centralized disclosure and extensive data quality controls, and samples of OGD from decentralized data disclosure (municipalities level), with no possibility of extensive quality controls as in the former case, hence with supposed lower quality.

Starting from measurements based on the quality framework, we were able to verify the difference in quality: the measures showed a few common acquired good practices and weaknesses, and a set of discriminating factors that pertain to the type of datasets and the overall approach. On the basis of this evaluation, we also provided technical and policy guidelines to overcome the weaknesses observed in the decentralized release policy, addressing specific quality aspects.

Keywords: Open Government Data; Open Data Quality; Government Information Quality; Data Quality Measurement; Empirical Assessment.

1. INTRODUCTION

Open data are data that "can be freely used, modified, and shared by anyone for any purpose" (Web ref. 1). Compared to proprietary frameworks, digital commons such as open data are characterized - from both a legal and a technical point of view - by lower restrictions applied to their circulation and reuse. This feature is supposed to ultimately foster collaboration, creativity and innovation (Hofmokl, 2010).

At all administrative levels, the public sector is one of the major producers and holders of information, which ranges, e.g., from maps to companies registers (Aichholzer and Burkert, 2004). During the last years, the amount and variety of open data released by public administrations across the world has been tangibly growing (see, e.g., the Open Data Census by the Open Knowledge Foundation (Web ref 2)), while increased political awareness on the subject has been translated in regulation, including the revised of the EU Directive on Public Sector Information reuse in 2013, as well as national roadmaps and technical guidelines. Releasing public sector information as open data can provide considerable added value, 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, also for commercial purposes (Vickery, 2011), and in general improve transparency of public institutions (Stiglitz et al., 2000; Ubaldi, 2013) and distributed ability to interpret complex phenomena (Janssen et al., 2012).

From the point of view of data reusers, we should take into account the role of the so-called infomediaries, i.e., players that are able to interpret data and present them effectively to the general public (Mayer-Schönberger & Zappia, 2011), with a highly diversified set of business models (Zuiderwijk et al., 2014). More in general, open data consumers manipulate data in many ways, ranging, e.g., from data integration to classification, also depending on the complementary assets they hold (Ferro & Osella, 2013). Considering this, legal and technical openness of datasets is not sufficient, by itself, to create a prolific reuse ecosystem (Helbig N. et al., 2012): failures in providing good quality information might impair not only the reuse of the data, but also the usage of the institutional portals (Detlor et al., 2014). Attempts to increase meaningfulness and reusability of public sector information also imply representing and exposing data so that they can be easily accessed, queried, processed and linked with other data with no restrictions (Sharon D.J., 2010). Although sharing common functionalities, the most used softwares for open data publication may adopt different approaches to ensure the above (Iemma, Morando & Osella, 2014).

On top of these considerations, it is necessary to consider that low-quality data provision increases the cost (in its wider meaning) of accessing and interpreting data: it is this cost and not the poor quality of Open Government Data (OGD) in itself that motivates the paper at hand. These additional costs imputable to low quality data depend on multiple factors, such as the nature of the data or the type of use and users (Kim, 2002). Such costs have been reported in the literature mainly for enterprises, affirming that high quality data are fundamental factors to a

company's success (Madnick et al., 2004; Haug et al., 2009; Even & Shankaranarayanan, 2009). This evidence has been recently reported also for Government Data (Zuiderwijk et al., 2015), whose disclosure policies have spread worldwide only after USA 2005 new guidelines on the Freedom of Information Act or after the presidency of the U.S.A. issued the 2013 Executive Order "Making Open and Machine Readable the New Default for Government Information". Even Italy, which was one of the early adopters of the "Open by Default" principle, made an explicit policy only 2012 (Decree No. 179). As a consequence, our logical argumentations are by construct undetermined by published data on the negative impact of bad OGD quality (see, for instance: (Web ref 3), (Web ref 4)). Therefore it is self evident that when low-quality data are released as open data, reuse will be discouraged, and/or several re-users will invest in checking and increasing such quality in a decentralized and uncoordinated way: understanding poorly documenting datasets and performing data cleansing activities represent a significant proportion of the effort necessary to reuse Open Data. This represents a waste of resources. Public administrations are the cheapest cost avoider with respect to the social waste represented by the poor quality of OGD, and they are supposed to be players with a stronger incentive in investing for the public good: increasing the quality of OGD, they could foster reuse and focus the resources of re-users on added value services. However, assessing the quality of OGD is (one of) the necessary preliminary step(s) to motivate public administrations to invest on improving OGD quality.

This motivates our study, together with the fact that re-users will directly benefit from a higher quality of OGD. In the paper at hand, we suggest that a widely adopted data quality model for open data, and a set of actionable metrics are necessary tools to achieve data quality improvement, as it has been recently advocated in the literature (Zuiderwijk et al., 2015) (Umbrich et al., 2015), contributing to unlock the potential of reuse. Formally, Data Quality is defined in the ISO 25012 Standard as "the capability of data to satisfy stated and implied needs when used under specified conditions"(ISO 25012). In addition to this definition, Data Quality is often defined in the literature as "fitness for use" (Wang and Strong, 1996) (Batini et al., 2009), i.e., the ability of a data collection to meet users' requirements: thus, "fitness for use" emphasizes the importance of taking a consumer viewpoint of quality.

In this context, we built an evaluation framework based on the analysis of the methodologies for data quality measurement documented in literature, and tailored according to users' needs. Afterwards, we used the resulting evaluation framework to compare the quality of two samples of OGD: in the first case, data are released in a decentralized way (i.e., with no common structure) by different local administrations; in the second, data structures are standardized at governmental level (both cases are samples from Italian Open Government Data). The comparison has been driven by the following research question: *is the application of a metric-based evaluation framework for OGD quality able to detect acquired good practices, weak aspects and discriminating factors between samples of OGD from two different disclosure strategies?*

The remainder of the paper is organized as follows. We provide motivations and comparison of our approach with related work in Section 2. In Section 3 we build the quality framework supported by a data quality model from the literature, afterwards we set up an initial set of metrics. Then, we apply the framework (Section 4) to observe differences in quality between OGD released at national level (obtained integrating data from different regions, with presumably higher data quality), and OGD at municipality level and therefore characterized by decentralized disclosure (and presumably lower data quality). We package the results (Section 5) in a set of acquired good practices, weaknesses to overcome and discriminating factors; in addition, we complement these results with a list of technical guidelines in terms of tools, processes and research directions (Section 6) to improve quality of data whenever a centralized disclosure of data is not possible due to costs and nature of data. We discuss the limitation of our approach in Section 7, and provide our roadmap and recommendations for future work in Section 8.

2. RELATED WORK

2.1 OGD quality problems reported by practitioners

Several studies provided examples that contribute to show how issues related to poor data quality can be widespread, and potentially hamper an efficient reuse of open data.

Allison (2010) reports problems of accuracy, aggregation and precision in Open Government Data, such as bad manual transposition of zip codes in public archives. Another example comes from the monthly reports of the American Nuclear Regulatory Commission, where spent fuel quantities are recorded: data were bouncing both ways from December 1982 to May 1983, while the trend for such type of data must have been only positive (Barlett & Steele, 1985).

Aggregation problems were reported on FedSpending.org, that keeps records on federal contract and grant data: data about companies that acquired new parent companies were wrongly aggregated, making impossible to track the money received after mergers or spin-offs. Another example of poor aggregation and insufficient details comes from a project by the Sunlight Foundation called Fortune 535, in which the organization used the personal financial disclosure forms that every U.S. Member of Congress is obliged to fill since 1978. Data were collected in ranges of income (e.g., from $1 to $1,000, from $1,001 and $15,000, etc.), but the Congress changed these ranges several times, making it virtually impossible to create consistent time series (e.g., to analyze which members of Congress accumulated richness during public service).

Other examples concern lack of integration. For instance, Tauberer (2012) reports that the two chambers of the U.S. Congress disclosed their data with completely different schema and IDs for Members of Congress, making data integration very difficult: as a consequence, merging or comparing data requires significant extra-effort, which could be easily avoided by achieving a better coordination between the data holders.

Sunlight's labs director Tom Lee reports data quality problems in a blog post entitled "How Not to Release Data" (Web ref 5). The data about White House e-mail records was released in form of printouts from the record management system (ARMS) and then transmitted via fax to Clinton library and re-digitized through OCR. At this point the document was encoded in PDF and released. The result was badly formatted, duplicative and missing information. Moreover, in a recent analysis performed by the authors of this paper on open datasets released by the City of Torino (Italy), problems regarding absence of metadata, not reported measuring system for geographical information and missing data (Web ref. 6), were identified. Information quality issues can affect also re-users in the academic field: for example Whitmore (2014) reported information quality issues (mainly inaccurate and incomplete data) as one of the barriers impairing the prediction capabilities of open data from US Department of Defense to predict future wars.

Finally, the Open Knowledge Foundation collected examples of 'bad' data provision, mainly from public data holders (Web ref. 7). Issues relate with data structuring, formats, and other poorly user-friendly practices.

2.2Existing approaches to OGD quality measurement

The issue of open data quality has been partially addressed in recent years. In 2006, Tim Berners-Lee published a deployment scheme for open data, based on five -- incremental and progressively demanding -- requirements represented as "stars" (Berners-Lee, 2006). A 5-stars open dataset should comply with all of these requirements:

1. Available on the web, any format provided data has an open license;
2. Available as machine-readable structured data (e.g. Excel instead of image scan);
3. Available non-proprietary format (e.g. CSV instead of Excel);
4. Make use of open standards from W3C (RDF and SPARQL) and URIs to identify things;
5. Link data to other providers' data to provide context.

Although widely cited, this schema proposed by Tim-Berners Lee covers only a specific quality aspect, i.e., the format or encoding used to publish the data. As a consequence, a dataset can reach the 5 stars level while showing at the same time poor quality. For instance, a common problem is data accuracy: for example typing or syllabling errors when data input is manual, or associating values to the wrong instances because of a software misbehavior. In addition, as we have seen in Section 2.1, the quality of open data does not concern accuracy only, but also other characteristics: such as completeness, consistency and timeliness (Catarci & Scannapieco, 2002). In the literature of data quality, several authors contributed to build an extensive list of data quality characteristics: see for instance Wand and Wang (1996), Wang and Strong (1996), Redman (1996), Jarke et al. (1995), Bovee et al. (2003), Naumann (2002), (Batini & Scannapieco, 2006). However one problem we noticed is that usually different definitions to same quality characteristics have been given. Another issue we found is that only a few assessments has been done so far, and the different definitions of the quality characteristic under study resulted in different metrics used to measure the same characteristics, as we are going to report herein.

Ubaldi (2013) proposed an analytical framework to assess of Open Government Data (OGD). The author developed a large set of metrics at very heterogeneous points of view (e.g., political, organizational, technical). Data quality is measured 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), re-use (e.g., number of apps developed with the data). Although this work is probably the very first comprehensive approach for the assessment of open data, from the technical point of view it remains quite at high level, because all metrics proposed are at portal level. In addition, no evaluation of the metrics is available.

A more detailed study was conducted by Maurino et al. (2014): the authors analyzed 50 datasets from Italian OGD at various administrative levels (regions, provinces, municipalities) in terms of completeness, accuracy and timeliness. Completeness is computed with respect to the availability and accessibility of the document through internal or external links, accuracy in terms of its format (using a 3 levels scale instead of the standard de facto 5-stars from (Berners-Lee, 2006)), while timeliness as presence or absence of updates. The measurements proposed by

Maurino et al. (2014) are at dataset level, but 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 evaluation shed a light on data quality problems of Italian 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.

Finally, regarding timeliness, Atz (2014) proposed the "tau metric", i.e. a metric that captures the percentage of datasets up-to-date in a data catalogue, and he applies it to three different portals (World Bank, the UK data catalogue and the London data store). The author computed the metric on the datasets and then aggregated to form a single indicator of Timeliness, which also discriminates 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 of Maurino et al (2014).

Our work takes inspiration and further motivations from these studies and tries to overcome their limitations. In particular we aimed to address the following gaps:

a) We observed that existing studies in literature lack reference to a comprehensive theoretical framework with univocal definitions of quality characteristics: we propose an evaluation framework build on top of a quality model adapted from the literature of data quality with univocal definitions of quality characteristics.
b) We observed that existing works in literature assess the quality of OGD mainly at portal level: we define a set of quantitative indicators, some of them specific to OGD, for a subset of the available quality characteristics and at the most granular level of measurement, which is cell level (according to tabular representation) or dataset level when otherwise not possible. We assess the suitability of the metrics by using them to compare a sample of OGD from two different disclosure strategies, to reveal their acquired good practices, weak aspects and discriminating factors.

The approach we propose is similar to the one adopted by Behkamal et al. (2014) for Open Linked Data quality: the authors investigated a set of quality characteristics taking as reference the ISO25012 standard data quality model. The authors build a set of 20 metrics related to semantic and syntactic accuracy, uniqueness, completeness and consistency. They verify the suitability of the proposed 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.

With respect to the approach of Behkamal et al. (2014), our work differentiates in terms of quality model and target data. The authors refer to a data quality model ("SQUARE" - ISO25012), which is a subset of the one we considered (SPDQM, i.e. SQUARE Aligned Portal Data Quality Model). In addition, their metrics are specific to linked data, which are in form of triples while our metrics apply to tabular data: Linked Data are only a fraction of the whole disclosed Government Data. More specifically, in Italy, public datasets in RDF format are 1975 while datasets in csv format (thus, tabular) are 6471 (Web ref 8).

In the following section we report how we select the theoretical framework and we operationalize it with metrics, comparing with the literature work.

3. OPEN GOVERNMENT DATA QUALITY MEASUREMENT FRAMEWORK

Our approach to define an open data quality measurement framework consists of three parts:

- identification of the most suitable data quality model as theoretical support of the measurement framework (subsection 3.1),
- methodology for the selection of data quality characteristics and metrics (subsection 3.2),
- results on the selection of data quality characteristics and metrics (subsection 3.3)

3.1 Data quality model selection

There are, available in the literature, several theoretical frameworks, intended as set of metrics, characteristics and dimensions, for assessing data quality (DQ), each of these being usually part of a data quality methodology. Such methodologies are typically defined as sequences of activities that can be divided into three phases (Batini et al., 2009):

1. State Reconstruction: which collects contextual information about the data, like organizational process and services, quality issues and corresponding costs, data collections and related procedures.
2. Assessment/measurement: which is aimed at measuring the quality of data with respect to some relevant quality dimensions.
3. Improvement: which includes all the steps and procedures for attaining higher data quality targets.

Our focus is on phase 2. Batini et al. (2009) collected from the literature an exhaustive list of data quality methodologies, that are reported in Table 1.

Table 1 Methodologies considered in (Batini et al., 2009)

Methodology Acronym	Extended Name
TDQM	Total Data Quality Management
DWQ	The Datawarehouse Quality Methodology
TIQM	Total Information Quality Management
AIMQ	A methodology for information quality assessment
CIHI	Canadian Institute for Health Information methodology
DQA	Data Quality Assessment
IQM	Information Quality Measurement
ISTAT	ISTAT methodology
AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality)
DaQuinCIS	Data Quality in Cooperative Information Systems
QAFD	Methodology for the Quality Assessment of Financial Data
CDQ	Comprehensive methodology for Data Quality management

This list does not contain a relevant model, which is the Square-Aligned Portal Data Quality Model (SPDQM, Moraga et al., 2009), because it was published shortly after the book of Batini et al. (2009). The SPDQM was built upon the Portal Data Quality Model (PDQM, Calero et al. (2008)) and the SQuaRE (ISO/IEC 25012, 2008) standard, both also not contained in the list of Table 1. SPDQM contains a set of 42 characteristics (30 characteristics from PDQM, 7 characteristics from SQUARE, the remaining 5 characteristics were added after a systematic literature review), which are organized in two viewpoints and four categories:

- Inherent
  - Intrinsic: This denotes that data have quality in their own right
- System dependent
  - Operational: The data must be accessible but secure
  - Contextual: Data Quality must be considered within the context of the task in hand
  - Representational: Data must be interpretable, easy to understand, concisely, and consistently represented

The most commonly reported quality problems (see also Section 2.1) correspond to either intrinsic properties of the data or properties depending on the context or the usage (Bovee et al., 2003). Since OGD typically span heterogeneous domains and they are subject to the most diverse usage from their consumers, our opinion is that it is preferable to select the dimensions that address the intrinsic aspects of data quality. In this viewpoint, SPDQM contains the most complete set of characteristics (12) when compared to the other models. In addition, SPDQM presents a set of basic characteristics shared by almost all the models (accuracy, completeness, timeliness), and provides characteristics as traceability, compliance and understandability, which are less considered by other frameworks, but are nonetheless important for OGD.

Thus, considering the above, we used the SPDQM as theoretical support for our framework.

3.2 Quality characteristics and metrics definition: methods

The next step is the selection of a subset of quality characteristics from SPDQM and the definition of metrics for those characteristics.

Concerning the selection of a sub-set of quality characteristics from the full set defined by SPDQM, we used the results from a survey previously conducted by the authors. The survey was conducted in 2013 among the participants of two hackatons aimed at reusing OGD. We collected answers from 15 developers. We focused on two items (out of 14) in the questionnaire used for that survey, the whole instrument and the measures are reported in Appendix B.1. The analyzed item ("Which problems did you find working with open data?", "Which aspects of data quality would you like to improve?") collects the issues as reported by practitioners. We built a complete list of the most common issues starting from the answers and mapped them onto the data quality characteristics of SPDQM. This approach, though suffering from a limited generalizability, is far less biased than a selection based solely on the personal believing of the research team.

Concerning the definition of the metrics, we relied on the principles of Kaiser et al. (2007):

- Measurability: the metrics should be normalized and at least interval scaled.
- Interpretability: the metrics have to be comprehensible. Their definition should have right amount of information in order to be interpretable.
- Aggregation: it should be possible to quantify data quality at an attribute level, as well as tuple, dataset or database level. In this way metrics would have a semantic consistency on all the levels. Moreover the metrics should permit value aggregation at a certain level in order to obtain the metric at a higher level.
- Feasibility: in order for the metrics to be applicable in a practical way, they should be based on determinable input parameters and be preferably automatable.

In addition to these indications, metrics can be classified as either objective, when they are based on quantitative metrics, or subjective, when they are based on qualitative evaluations from users (like in surveys). In this work we will give more emphasis to quantitative measures: quantitative indicators are important for triangulation with assessments based on qualitative measures, like questionnaires or experts' opinions, which suffer from subjectivism and inconsistency. With such a list of desired requirements at hand, we searched the literature for metrics on the selected SPDQM characteristics that satisfy them, and when no metric was found, we formulated it ex-novo. Also, when possible this was done taking as unit of measure the cell of the dataset, when not, the metrics are at dataset level. We took into account both the definition of quality characteristic and, whenever possible, the type of problem reported by developers.

3.3 Quality characteristics and metrics definition: results

Table 2 summarizes the type of problems emerging from the survey and links them to the data quality characteristics of SPDQM. The mapping was achieved by comparing the definition of the characteristics with the issues highlighted by developers. The classification was agreed upon in a meeting involving four of the authors of this work.

Table 2 .Problems found in the exploratory survey

Problem found in survey	Related quality characteristic	Intrinsic	System-dependent
Incomplete data	Completeness	X	
Format not compliant to well known standard	Compliance		
Lack of data source traceability	Traceability	X	X
Incongruent data	Consistency	X	
Out-of-date data	Expiration, Currentness	X	
Lack of Metadata	Compliance, Understandability		
Errors	Accuracy	X	
High time to understand data	Understandability	X	X

Part of the mapping was straightforward: the issues "Incomplete data", "Lack of data source traceability", "Incongruent data", "Errors" and "High time to understand data" fit very well to quality characteristics of SPDQM. For the other mappings a few further words have to be spent. The issue "Out-of-date data" could refer both to time validity and data obsolescence, for this reason in Table 2 refers to both expiration and currentness. Some discussion has to be done for issue "Lack of Metadata". Answers to the questionnaire showed that developers encountered understandability problems. In our theory, this happens due to poor metadata that do not provide useful guidance. Although we could not test this cause-effect relationship, we believe that it is safe and reasonable to map the code "Lack of Metadata" with Understandability, given that also in the literature metadata are considered fundamental for the right comprehension of the dataset (see Reiche & Hofig (2013)). In addition, "Lack of Metadata" is also mapped to compliance due to the existence of a standard for metadata in Open Government Data sets. Finally, the issue "Format not compliant to well known standard" had no clear corresponding quality characteristic: we also mapped it to compliance and we measured it with the compliance to the 5 Stars open data format scheme from Tim Berners-Lee (2006).

Table 3 contains the metrics we defined for each of the selected quality attributes, reporting name and descriptions. The formulas used to compute them and the literature references are shown in the Appendix A, while specific tuning of the implementation for the datasets analyzed are provided later in Table 5, when the data sources are also presented.

Table 3. Metrics definitions and description

Characteristic	Metric	Level	Description
Traceability	Track of creation	Dataset	Indicates the presence or absence of metadata associated with the process of creation of a dataset.
	Track of updates	Dataset	Indicates the existence or absence of metadata associated with the updates done to a dataset.
Currentness	Percentage of current rows	Cell	Indicates the percentage of rows of a dataset that have current values, it means that they don't have any value that refers to a previous or a following period of time.
	Delay in publication	Dataset	Indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the dataset) and the period of time referred by the dataset (week, month, year).
Expiration	Delay after expiration	Dataset	Indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, year).
Completeness	Percentage of complete cells	Cell	Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column).
	Percentage of complete rows	Cell	Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell.
Compliance	Percentage of standardized columns	Cell	Indicates the percentage of standardized columns in a dataset. It just considers the columns that represent some kind of information that has standards associated with it (i.e. geographic information).
	eGMS Compliance	Dataset	Indicates the degree to which a dataset follows the e-GMS standard (as far as the basic elements are concerned, it essentially boils down to a specification of which Dublin Core metadata should be supplied)
	Five star Open Data	Dataset	Indicates the level of the 5 Star Open Data model in which the dataset is and the advantage offered by this reason.
Understandability	Percentage of columns with metadata	Cell	Indicates the percentage of columns in a dataset that have associated descriptive metadata. This metadata is important because it allows to easily understanding the information of the data and the way it is represented.
	Percentage of columns in comprehensible format	Cell	Indicates the percentage of columns in a dataset that are represented in a format that can be easily understood by the users and it is also machine-readable.
Accuracy	Percentage of accurate cells	Cell	Indicates the percentage cells in a dataset that have correct values according to the domain and the type of information of the dataset.
	Accuracy in aggregation	Cell	Indicates the ratio between the error in aggregation and the scale of data representation. This metric only applies for the datasets that have aggregation columns or when there are two or more datasets referring to the same information but in a different granularity level.

4. Application of the framework: material and methods.

4.1 Goals and object of the study

Our aim is to use the previously defined evaluation framework to quantitatively assess the quality of Open Government Data. To achieve our goal, we needed an oracle for OGD of high quality. For this reason we selected an internationally recognized good example of OGD, i.e. the Open Coesione project, which has been ranked 4th at the 2014 Open Government Awards (Web ref. 9). We chose Open Coesione and not the first classified for two reasons: it was the representative project for Italy and we had a direct contact with its managers at the Italian Ministry General State Accounting Department of Italy's Ministry of Economy and Finance. Open Coesione data has been disclosed with a centralized release strategy with extensive quality controls. We compared Open Coesione with samples of OGD from municipalities and therefore decentralized data disclosure, with no possibility of extensive quality controls as in the former case, hence with supposed lower quality. In fact, considering the workflow leading to the publication of open data, a difference in the quality of the published data may be due to a difference in the quality of the original datasets or to a difference in the publication pipeline. The original datasets object of the study, are in both cases released by municipalities: therefore, it is safe to assume that the quality of data at the source is comparable. However, in the OpenCoesione case, while data are aggregated at regional level, they also undergo a quality improvement process (see Section 4.2 for details): therefore, we can assume that if any quality difference is measured between the two samples, it is due to the publication pipeline.

In this way, we are able to understand the suitability and limitations of the metrics in our framework to correctly identify the quality differences. We formalize our goal with the Goal Question Metric template (Basili et al., 1994):

Object	Open Government Datasets
Purpose	Understand acquired good practices, weak aspects, discriminating factors
Focus	Intrinsic data quality
Stakeholder	Data releaser, data user
Context	Open Government Data

Our resulting research question is:

Is the application of a metric-based evaluation framework for OGD quality able to detect acquired good practices, weak aspects and discriminant factors between samples of OGD from two different disclosure strategies?

4.2 Datasets analyzed

Open Coesione contains data about the fulfillment of the investments and related projects by the Italian central government and regions using the 2007-2013 European Cohesion funds. At the time of this writing, the portal contains data about 900.000 projects, worth 90 billion Euros of financings. Data and information about territorial cohesion policies concern projects fundamentals, amount of funding, locations, involved subjects and completion times. The information about the 2007-2013 funds is gathered by a single monitoring system managed by the General State Accounting Department (RGS – Ragioneria Generale dello Stato) of Italy's Ministry of Economy and Finance. The monitoring system entails a data integration process and a data quality control process to homogenize the information coming from the different Italian administrative regions. The data quality control process concerns mainly the identification codes of the projects and the financial and time variables: checks are done on data completeness and correctness, and on the consistency between different variables. In particular, regarding consistency, there are 30 controls, which could determine the deletion of a project from the dataset, and 14 checks that arise a warning: a sample control on six funding programs and six regions resulting in more than 70K projects has generated about 18K drop outs and about 110K warnings. Most of the problems were related to inconsistencies of projects identifiers in different datasets, missing data, payments that overcome the available funding, inconsistent addresses.

The specific dataset object of this study contains the projects funded in the European Social Fund 2007/2013 (Web ref. 10). We used the snapshot of 31st December 2013, which contains about 55 million observations of 75 variables (approximately 1 GB size). Variables concerns: projects identification data; projects thematic classification according to Italian and European schemas; funding programs; financial data (funding and payments); dates; control variables used in the quality checks. After discussing with the stakeholders at the RGS, we focused our analysis only on the variables not included in their quality checks procedure. This subset of variables concerns the amounts of financial support by the different institutions (European Union, Italian Government, Italian local governments) and project dates: in total 22 variables (i.e. columns) and about 16 million data-points.

Concerning the decentralized data disclosure, we needed more than one dataset to reduce as much as possible noise and bias in the data sources, however on comparable topics. We decided to select our sample from regional administrative municipalities because at regional level the published data were too diverse and not comparable. Therefore we searched through the single municipalities portals for catalogues containing data about the same topic (such as resident citizens, hospitals, etc.). However, even at that level data were very diverse and we ended up our search with datasets about three topics: resident citizens, marriages, and business activities. Table 4 shows which dataset type was found in which city portal. Details and URLs for each datasets are reported in Appendix B.2.

Table 4. Datasets used with decentralized disclosure

Datasets	Torino	Roma	Milano	Firenze	Bologna
Residents	X	X	X	X	X
Marriages	X		X	X	
Business	X	X	X		

4.3 Analysis methodology

Automatic computation of metrics has been possible for completeness (percentage of complete cells, percentage of complete rows) and accuracy (percentage of accurate cells, accuracy in aggregation). For all other quality dimensions and related metrics, manual computation was necessary. All measures have been normalized to the interval [0,1] (see Appendix A for details).

In certain cases metrics were not applicable or undefined. A metric is considered not available (NA) if it measures a characteristic that does not apply to the dataset under study: for example, the percentage of columns adhering to a standard can be NA if the type of data contained in that column is not regulated by standards. While, a metric is considered not defined (ND) when there is not enough information to compute it: for example, the delay in publication is not defined when the publication date is missing either on the web site or within the metadata. We considered ND data equivalent to 0.

Another special case concerns empty cells in the Open Coesione dataset. Empty cells could be considered either as belonging to the domain or not: as a matter of fact, an empty cell in column "obtained funding" could be interpreted as a zero because the project didn't obtain any funding yet. After a preliminary analysis of the data in conjunction with the data releasers, we decided to compute completeness metrics including the empty cells in the domain of validity, so an empty cell still contribute to computation of the metric (see formula Appendix A).

As we have seen, although the metrics defined are applicable to any dataset in tabular form, their implementation might need small but necessary tailoring to be applied to specific data domains. This is not a limitation of the metrics but rather a peculiarity of OGD, whose domains vary a lot and data are heterogeneous. We summarize in Table 5 the dataset-specific refinements that were necessary and the reasons.

As far as the analysis of results is concerned, for each defined metric m_i, we built a pair of null/alternative hypotheses:

$H_0 : m_{i,\text{OPENCOESIONE}} = m_{i,\text{ MUNICPALITIES}}$

$H_A : m_{i,\text{OPENCOESIONE}} \neq m_{i,\text{ MUNICPALITIES}}$

The null hypothesis is that there is no difference between the Open Coesione datasets and the sample of municipalities' datasets. We perform the two tailed Mann Whitney test (Sachs, L. (1982)) with standard confidence level of 95% to compare the measurements on the two groups (for each metric). As a consequence, if p-value is less then 0.05, the null hypothesis is rejected in favor of the alternative one.

Due to the low number of datasets, we are aware that the test could be conservative and only looking at the p-value could be misleading in case of small differences. For this reason we report also confidence intervals. Additionally, in order to better adhere to our stated goals (see Section 4.1), we also run a more pragmatic analysis methodology and classify the metrics into three categories:

- Acquired good practices, in relation to metrics that in all datasets of the analyzed regions/municipalities have measures ≥ 0.50
- Quality aspects to be improved, in relation to metrics that in all datasets of the analyzed regions/municipalities have measures < 0.50
- Discriminating factors, i.e. metrics whose measures change significantly across the various data sources and determine in which quality aspects a dataset is different from the others.

The choice of threshold 0.50 is based on the fact that the resulting indicators are polarized (see Figure 1) and therefore such a threshold is reliable.

Table 5. Metrics refinement

Metric	Dataset-specific refinement	Reason
Percentage of syntactically accurate cells	The domain of values may vary from attribute to attribute. Every dataset needs to have specified domains for each attribute	Not all the attributes have the same domain
Percentage of complete cells	If there is no information in the metadata, some assumptions must be made about the interpretation of null values and (possible) default values (e.g. date of birth 1/1/1900).	Default values may be missing values, while for null values it must be checked with stakeholder if they are admitted in the domain of the attribute (e.g., in the case of optional values).
Percentage of complete rows	The same as in "Percentage of complete cells"	The same as in "Percentage of complete cells"

5. RESULTS

We report the measurements obtained in Figure 1. The first three graphs correspond to the municipality-level datasets, the last one to Open Coesione with data aggregated from all regions. Not Available data in municipalities are reported as an empty cell in the corresponding embedded table (and do not contribute to the evaluation), while Not Defined data is equivalent to 0 (and they do contribute to the evaluation) as discussed in Section 4.3.

We report in Table 6 the average of the measurements and the standard deviations. The comparison between Open Coesione and the sample of municipality-data is summarized in Table 7 according to the analysis methodology described in Section 4.3: a "+" indicates an acquired good practice, a "-" indicates an aspect to improved, a "+/-" a discriminating factor among the regions/municipalities.

We observe that most of the variability resides in understandability. Differences are also in completeness (the only aspect in favor of municipalities), accuracy, traceability, currentness and expiration. Finally, there are equal levels of traceability, compliance and expiration.

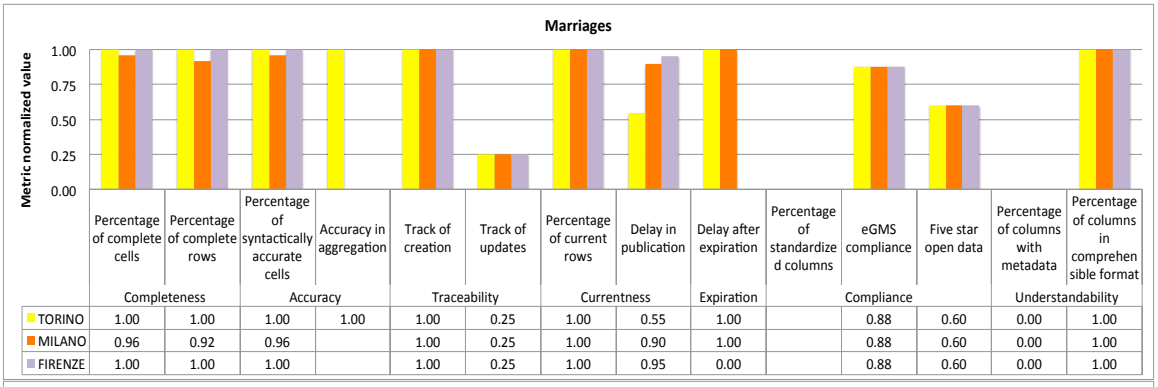
Last two columns of Table 7 show the p-value of the Mann-Whitney test (in bold when statistically significant) and the confidence interval of the difference between the two groups. Statistically significant difference is found in the following metrics: Percentage of complete rows, Percentage of syntactically accurate, Track of updates, Delay in publication, Delay after expiration, eGMS compliance, Percentage of columns with metadata. We observe a border value for Percentage of columns in comprehensible format. The confidence intervals reveal that except for Percentage of complete rows, all statistically significant differences are in favor of the Open Coesione datasets.

Table 6. Metrics computation : descriptive statistics

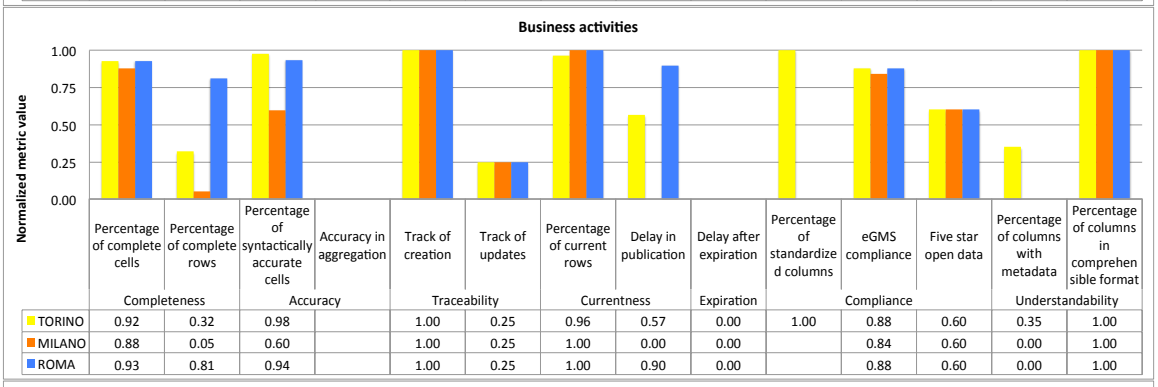
		Open Coesione (n=20)		Municipalities (n=11)	
Dimension	**Metric**	**Average**	**Standard deviation**	**Average**	**Standard deviation**
Completeness	Percentage of complete cells	0.94	0.07	0.92	0.15
	Percentage of complete rows	0.00	0.00	0.74	0.38
Accuracy	Percentage of syntactically accurate cells	1.00	0.00	0.91	0.18
	Accuracy in aggregation	1.00	0.00	1.00	0.00
Traceability	Track of creation	1.00	0.00	1.00	0.00
	Track of updates	0,50	0.00	0.23	0.08
Currentness	Percentage of current rows	1.00	0.00	1.00	0.01
	Delay in publication	0.97	0.01	0.59	0.35
Expiration	Delay after expiration	0.99	0.00	0.26	0.45
Compliance	Percentage of standardized columns	1.00	0.00	1.00	NA
	eGMS compliance	0.92	0.00	0.88	0.01
	Five star Open Data	0.60	0.00	0.60	0.00
Understandability	Percentage of columns with metadata	1.00	0.00	0.03	0.11
	Percentage of columns in comprehensible format	1.00	0.00	0.83	0.39

Table 7. Comparison of the evaluations

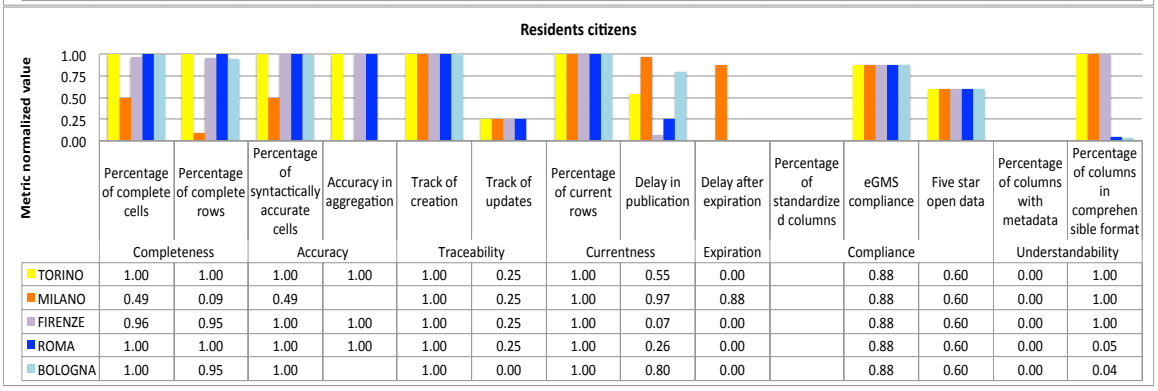
Dimension	**Metric**	**Open Coesione**	**Municipalities**	**P value**	**Conf. interval**
Completeness	Percentage of complete cells	+	+	0.55	{-0.04; 0.04}
	Percentage of complete rows	-	+/-	**< 0.05**	{-1.00;-0.81}
Accuracy	Percentage of syntactically accurate cells	+	+/-	**< 0.05**	{-0.001; 0.04}
	Accuracy in aggregation	+	+	NaN	NaN
Traceability	Track of creation	+	+	NaN	NaN
	Track of updates	-	-	**< 0.05**	{0.25; 0.25}
Currentness	Percentage of current rows	+	+	0.20	{0 ; 0}
	Delay in publication	+	+/-	**< 0.05**	{0.08; 0.43}
Expiration	Delay after expiration	+	+/-	< 0.05	{0.99; 0.99}
Compliance	Percentage of standardized columns	+	+	NaN	NaN
	eGMS compliance	+	+	**< 0.05**	{0.04; 0.04}
	Five star Open Data	+	+	NaN	NaN
Understandability	Percentage of columns with metadata	+	-	**< 0.05**	{1.00; 1.00}
	Percentage of columns in comprehensible format	+	+/-	0.06	{0 ; 0}



	Percentage of complete cells	Percentage of complete rows	Percentage of syntactically accurate cells	Accuracy in aggregation	Track of creation	Track of updates	Percentage of current rows	Delay in publication	Delay after expiration	Percentage of standardized columns	eGMS compliance	Five star open data	Percentage of columns with metadata	Percentage of columns in comprehensible format
	Completeness		Accuracy		Traceability		Currentness		Expiration		Compliance		Understandability	
TORINO	1.00	1.00	1.00	1.00	1.00	0.25	1.00	0.55	1.00		0.88	0.60	0.00	1.00
MILANO	0.96	0.92	0.96		1.00	0.25	1.00	0.90	1.00		0.88	0.60	0.00	1.00
FIRENZE	1.00	1.00	1.00		1.00	0.25	1.00	0.95	0.00		0.88	0.60	0.00	1.00



	Percentage of complete cells	Percentage of complete rows	Percentage of syntactically accurate cells	Accuracy in aggregation	Track of creation	Track of updates	Percentage of current rows	Delay in publication	Delay after expiration	Percentage of standardized columns	eGMS compliance	Five star open data	Percentage of columns with metadata	Percentage of columns in comprehensible format
	Completeness		Accuracy		Traceability		Currentness		Expiration		Compliance		Understandability	
TORINO	0.92	0.32	0.98		1.00	0.25	0.96	0.57	0.00	1.00	0.88	0.60	0.35	1.00
MILANO	0.88	0.05	0.60		1.00	0.25	1.00	0.00	0.00		0.84	0.60	0.00	1.00
ROMA	0.93	0.81	0.94		1.00	0.25	1.00	0.90	0.00		0.88	0.60	0.00	1.00



	Percentage of complete cells	Percentage of complete rows	Percentage of syntactically accurate cells	Accuracy in aggregation	Track of creation	Track of updates	Percentage of current rows	Delay in publication	Delay after expiration	Percentage of standardized columns	eGMS compliance	Five star open data	Percentage of columns with metadata	Percentage of columns in comprehensible format
	Completeness		Accuracy		Traceability		Currentness		Expiration		Compliance		Understandability	
TORINO	1.00	1.00	1.00	1.00	1.00	0.25	1.00	0.55	0.00		0.88	0.60	0.00	1.00
MILANO	0.49	0.09	0.49		1.00	0.25	1.00	0.97	0.88		0.88	0.60	0.00	1.00
FIRENZE	0.96	0.95	1.00	1.00	1.00	0.25	1.00	0.07	0.00		0.88	0.60	0.00	1.00
ROMA	1.00	1.00	1.00	1.00	1.00	0.25	1.00	0.26	0.00		0.88	0.60	0.00	0.05
BOLOGNA	1.00	0.95	1.00		1.00	0.00	1.00	0.80	0.00		0.88	0.60	0.00	0.04

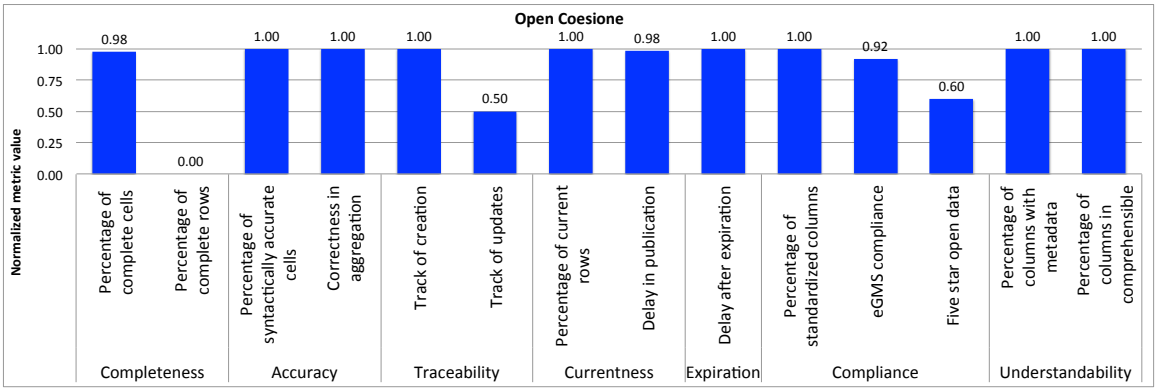


Figure 1. Computation of metrics (top three municipality level, on the bottom national level)

6. DISCUSSION

The measurements in Figure 1 and their summary in Table 6 and Table 7 lead us to the conclusion that the centralized data release, at national level, resulted in better quality than the decentralized one (at municipality level). Although this is expected due to the presence of a quality control process (described in Section 4.2), this gives us confidence on the capability of the framework to capture such quality difference. In addition, by applying the specific metrics defined, we are able to understand more precisely in what the instances of the two release strategies differed in terms of data quality.

In light of these measurements, we interpret the obtained results for the specific quality characteristics in sub-section 6.1. Then in sub-section 6.2, by following a line of reasoning by experience and analogy with the state of the art, we reason on possible tools and processes, which could be adopted to improve the quality of OGD on certain quality characteristics. This is useful for all those cases where setting up a data quality system is economically not convenient (e.g., data is too specific, resources are too scarce).

6.1 Comments on results

Understandability. Metadata at the lower administration level is an aspect that needs to be improved, because only in two cases (i.e., Torino/business activities and Firenze/marriages) we found metadata associated to the raw data, and in those two cases all the columns have metadata associated. Regarding comprehensible format, some cities performed very well, other obtained a metric score very close to zero: this was the case, for instance, where unit of measures were not specified.

Time aspects: Currentness and Expiration. The metrics on the time aspects were higher in Open Coesione, mainly because data are published at regular intervals of two months. On the contrary, on the municipality level, the assessment shows that data is generally current, although it is not always released as soon as it is available. We also observed many cases where expiration of data was not defined.

Accuracy. The data integration and quality checks in Open Coesione have a clear effect on the metric that measure syntactic accuracy. In the municipalities values were sometimes inconsistent with their domains. For instance, the business activities dataset in Torino had the following domain for column COD_COMP: {CFE, CF, E, EP}; however values in the corresponding column were in the following domain: {AE, CF, EP}. Notably the code AE is not present in the metadata schema.

Completeness. This metric is the only one where the centralized dataset performed worst. There are two possible explanations: either the integration of data with different structure results in empty cells for certain fields, which are not present in all regional sources, or this is due to the type of data stored in Open Coesione. By discussing with the stakeholders it was possible to

understand that most of the empty cells were due to not started projects, so the second explanation is more reasonable.

Traceability. In both cases the dataset creation information is always present, however there is no information available on data modifications and updates. In Open Coesione, in particular, although updates are made every two months, this does not guarantee that in the new release data from all regions are actually updated. This is the only dimension with a negative assessment in both types of datasets.

Compliance. Both samples obtained same positive results in compliance, which means that every time a specific standard was present, it was respected. Moreover, all datasets analyzed are at level three of Tim-Berners Lee scale (i.e., available, structured, in non-proprietary format). This is the most recurrent situation in Italian OGD (Web ref. 11): in fact, 8418 datasets out of 12084 are at level three at the time this article has been written. The rest of the distribution is the following: none at level one, 72 at level two, 2967 at level four, 627 at level five.

6.2 Guidelines for improving poor quality characteristics in OGD

Time aspects and Traceability: A possible solution to the problems illustrated above consists in applying versioning systems to open data, so that it is possible to easily access and compare different versions of the same data. Some proposals already exist in the literature. Rufus Pollock, founder and co-Director of the Open Knowledge Foundation (OKF), addresses this problem in a blog post (Web ref 12) and explores a solution based on well-known tools. The proposed solution is based on a data pattern made up of two pre-conditions: 1) data must be stored as line-oriented text, and more specifically as CSV file, 2) data must be stored on a GIT or Mercurial repository, which offer diff and merge tools. Canova et al. (2015) compare different approaches for versioning linked data. Sande et al. (2013) have explored a prototypal solution of versioning Linked Data with GIT, while *dat project* (Web ref. 13) offers a working alfa version for versioning CSV/XML/JSON files. Van de Sompel et al. (2010) explore an HTTP-based data versioning model applied to linked data, using a demonstrator built for DBpedia, also discussing how time-series analysis across versions of Linked Data descriptions could be elaborated.
These proposals rely on the assumption that versioning data improves traceability and encourages collaboration, as happened with open source software development.

Accuracy: *A* feedback mechanism to take note of the errors found by the users might be useful, perhaps in conjunction with the versioning system. However such practice could give rise to several issues, such as: how to clearly label and distinguish official and unofficial versions of the same dataset? How to manage (and fund the management of) this feedback channel? Assuming that a versioning system assists the process from the technical point of view: who has the rights to modify data? How is the process of re-validation managed? How much resources are required to supervise the users' feedback mechanism? Alexopoulos et al. (2014) discuss some of these

aspects and an open data infrastructure that provides feedback loops with all stakeholders. Kuk and Davies (2011) analyse the value chain of open government data via a multimethod study of the open data hackers in the UK.

Completeness. Completeness of data depends on both the type of data and the domain. For sure it is of fundamental importance declaring the domain of data (see also Understandability). In addition, simple instruments for data cleaning (e.g., Google Refine (Web ref. 14), Data Cleaner (Web ref. 15)) can be useful to assess the percentage of empty cells and understand the causes.

Understandability and Compliance. The first step to improve data understandability is to provide also metadata. We used the eGMS standard as a reference for the most important information to embed in metadata. A more comprehensive checklist is provided on the Opquast website for web quality (Web ref. 16). In addition, the Open Data Foundation provides also useful information on state of the art and best practices (Web ref. 17).

7. Limitations

This study is a first and partial attempt towards objective, reproducible, and scientifically based quality assessment of disclosed government data. It has some limitations, we discuss here the two that we consider the most important ones.

Low generalization of results. Due to time and effort required by manual evaluations of some metrics, and due to the difficulty of finding comparable datasets even in a large catalog as the Italian OGD portal, the number of datasets evaluated is small and results cannot generalized. However, finding a large number of datasets that represent the heterogeneity of the OGD universe and are still comparable (for instance in terms of domains) is a task that we believe being not feasible. For this reason we believe that the assessment framework proposed here has still high relevance, because it allows an evidence-based evaluation of Open Government Datasets. Replications from third parties of our analyses on different data sources will increase in the long run the generalizability of the obtained results.

Selection refinement of the quality characteristics. The selection of the quality characteristics was based on the results of a survey with a low number of respondents (n=15). For this reason we do not aim at considering those problems as representative of the most important quality characteristics for OGD from users and releasers perspectives. However, the results of the survey served us as a basis to prioritize the characteristics of intrinsic data quality that we wanted to operationalize first with metrics.

8. Conclusions and Future Work

Current approaches in literature to the problem of OGD quality lack of a comprehensive theoretical framework. In addition, most of the evaluations concentrate on open data platforms, rather than on datasets. In this work, we address these two limitations and provided a measurement framework to quantitatively assess the quality of OGD.

We tested the suitability of the framework by applying it on OGD samples from two different disclosure strategies: centralized, i.e. monitored at national level with data aggregated from several regional sources and in presence of a data quality process, and decentralized, without any type of orchestration and of presumably lower quality, selecting a set of municipalities' data in the same domains. We observed both common patterns and differences between the two compared release strategies. The metrics were able to show the benefits of the centralized data disclosure used as example of good quality OGD, as well as to quality issues originated from the samples of decentralized data disclosure that we analyzed. We also provided guidelines and references to improve Open Government Data on specific quality aspects, which might be valuable for those administrations which are not in the position to systematically apply a data quality process due to the relatively high costs associated, and as useful indications for future research as well.

Further ongoing work is devoted to understand whether the problems revealed by the metrics are able to predict problems experienced by developers when reusing the data. Finally, future work will focus on making the framework also applicable to non-tabular data and to define metrics for additional intrinsic quality characteristics. For instance the characteristics and relevant metrics chosen for the framework are not able to detect redundant and duplicate values in the datasets, or correctness of specific data formats. Also, the inability to make assessments in terms of some characteristics such as Currentness due to some metrics not being calculable only with the dataset at hand hurts the applicability of the framework and might require modification.

The long term goal of this study is to bring a data quality framework to a level where it can be turned into a tool that automatically assesses the quality of a dataset in terms of different characteristics, so that the negative aspects can be strengthened before the set is released to the public.

Acknowledgments

We would like to sincerely thank Aline Pennisi at Italian Ministry General State Accounting Department (RGS – Ragioneria Generale dello Stato) of Italy's Ministry of Economy and Finance. We also thank Giuseppe Procaccianti from VU University, Amsterdam, for his initial contribution to this work.

REFERENCES

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

Alexopoulos, C., Loukis, A., & Charalabidis, Y. (2014), *A Platform for Closing the Open Data Feedback Loop based on Web2. 0 functionality,* Journal of democracy and & Open Government, Vol.6, No 1.

Allison, B. (2010). *My Data Can't Tell You That*. In D. Lathrop, & L. Ruma, *Open Government – Collaboration, Trasparency, and Participation in Practice* (pp. 257-265). O'Reilly Media, Inc.

Atz, U. (2014) *The Tau of Data: A New Metric to Assess the Timeliness of Data in Catalogues*. In Proceedings of the International Conference for E-Democracy and Open Government (CeDEM2014), Krems, Austria.

Ballou, D. P., Wang, R. Y., Pazer, H., & Tayi, G. K. (1998). *Modeling Information Manufacturing Systems to Determine Information Product Quality*. Management Science, 44(4).

Barlett, D. L., & Steele, J. B. (1985). *Forevermore: nuclear waste in America*.

Basili, V. , Caldiera, G., & Rombach, H. D. (1994). *The Goal Question Metric Approach* .In: Encyclopedia of Software Engineering. John Wiley & Sons, P. 528–532.

Batini, C., & Scannapieco, M. (2006). *Data Quality, Concepts, Methodologies and Techniques*. Berlin, Springer.

Batini, C., Cappiello, C., Francalanci, C., & Maurino A.(2009). *Methodologies for data quality assessment and improvement*. ACM Comput. Surv. 41, 3, Article 16 (July 2009), 52 pages

Behshid Behkamal, Mohsen Kahani, Ebrahim Bagheri, and Zoran Jeremic. (2014). A metrics-driven approach for quality assessment of linked open data. J. Theor. Appl. Electron. Commer. Res. 9, 2 (May 2014), 64-79.

Bovee, M., Srivastava, R., & Mak, B. (2003). *Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality*. International Journal of Intelligent Systems, pp. 51-74.

Berners-Lee, T. (2006) *Linked data-design issues.* Tech. rep., W3C, http://www.w3.org/DesignIssues/LinkedData.html.

Calero, C., Caro, A., & Piattini, M. (2008). An applicable data quality model for web portal data consumers. *World Wide Web , 11* (4), 465-484.

Canova, L., Basso, S., Iemma, R. & Morando, F. "Collaborative Open Data versioning: a pragmatic approach using linked data". In International Conference for E-Democracy and Open Government 2015 (CeDEM15), Krems, Austria

Catarci, T., & Scannapieco, M. (2002) *Data Quality under the Computer Science Perspective.* Archivi Computer, Volume 2.

Detlor, B., Hupfer, Maureen E, Ruhi, U. & Zhao, L. (2013) *Information quality and community municipal portal use*, Government Information Quarterly, Volume 30, Issue 1, January 2013, Pages 23-32, ISSN 0740-624X, http://dx.doi.org/10.1016/j.giq.2012.08.00

Directive 2013/37/EU, available at: http://bit.ly/1mPZcLV

English, L. (1999.) *Improving Data Warehouse and Business Information Quality*. Wiley & Sons.

Even, A., & Shankaranarayanan, G. (2009). Utility cost perspectives in data quality management. Journal of Computer Information Systems, 50(2), 127-135

Ferro, E., & Osella, M. (2013). *Eight Business Model Archetypes for PSI Re-Use*. In "Open Data on the Web" Workshop, Google Campus, London.

Haug, A., Pedersen, A., & Arlbjørn, J.S. (2009). *A classification model of ERP system data quality.* Industrial Management & Data Systems, 109(8), 1053-1068. doi:10.1108/02635570910991292

Heinrich, B. (2002), *Datenqualitätsmanagement* in Data Warehouse-Systemen, doctoral thesis, Oldenburg.

Heinrich, B., Klier, M., and M. Kaiser. (2009). *A Procedure to Develop Metrics for Currency and its Application in CRM*. J. Data and Information Quality 1, 1, Article 5

Helbig N., Nakashima M., and Dawe Sharon S. (2012), *Understanding the Value and Limits of Government Information in Policy Informatics: A Preliminary Exploration* , Proceedings of the 13th Annual International Conference on Digital Government Research (dg.o2012) , June 4-7, 2012

Hofmokl, J. (2010). *The Internet commons: toward an eclectic theoretical framework*. International Journal of the Commons *, 4* (1), 226-250.

Iemma, R., Morando, F., & Osella, M. (2014). Breaking Public Administrations' Data Silos. eJournal of eDemocracy & Open Government, 6(2)

ISO/IEC. (2008). *25012 International Standard: Systems and software engineering - Software Product Quality Requirements and Evaluation (SQuaRE)-Data quality model.* Tech. rep., ISO/IEC.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). *Benefits, adoption barriers and myths of open data and open government*. Information Systems Management, 29(4), 258-268.

Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P., Eds. 1995. *Fundamentals of Data Warehouses*. Springer Verlag.

Jeusfeld, M.,Quix, C., & Jarke,M. (1998). *Design and analysis of quality information for datawarehouses*. In Proceedings of the 17th International Conference on Conceptual Modeling

Kaiser, M., Klier, M., & Heinrich, B. (2007), "How to Measure Data Quality? - A Metric-Based Approach" (2007). ICIS 2007 Proceedings. Paper 108.

Kim, W. (2002). *On Three Major Holes in Data Warehousing Today*. Journal of Object Technology, 1(4), 39-47. doi:10.5381/jot.2002.1.4.c3

Kuk, G., & Davies, T. (2011). *The Roles of Agency and Artifacts in Assembling Open Data Complementarities*. In Thirty Second International Conference on Information Systems.

Madnick, S., Wang, R., & Xian, X. (2004). *The design and implementation of a corporate householding knowledge processor to improve data quality*. Journal of Management Information Systems, 20(1), 41-49.

Maurino A., Spahiu B., Batini C., & Viscusi G. (2014). *Compliance with Open Government Data Policies: an empirical evaluation of Italian local public administrations*, Twenty Second European Conference on Information Systems, Tel Aviv 2014

Mayer-Schönberger, V., & Zappia, Z. (2011). *Participation and Power: Intermediaries of Open Data*. In 1st Berlin Symposium on Internet and Society.

Moraga, C., Moraga, M., Calero, C., & Caro, A. (2009). *SQuaRE-aligned data quality model for web portals*. Quality Software, 2009. QSIC'09. 9th International Conference on, (pp. 117-122).

Naumann, F.. 2002. *Quality-driven query answering for integrated information systems*. Lecture Notes in Computer Science, vol. 2261.

Redman, T. 1996. *Data Quality for the Information Age*. Artech House.

Reiche, K., & Hofig, E. (2013). *Implementation of Metadata Quality Metrics and Application on Public Government Data*. Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual (p. 236-241).

Sachs, L. (1982) *Applied Statistics. A Handbook of Techniques*. Springer-Verlag, New York — Heidelberg — Berlin 1982, 734 pp., 59 figs., DM 118,-

Sande, M.V., Dimou, A., Colpaert, P., Mannens, E.., & Van de Walle R. (2013). Linked Data as enabler for Open Data Ecosystems. Open Data on the Web 23 - 24 April 2013, Campus London, Shoreditch.

Sharon Dawes S. (2010). Stewardship and Usefulness: Policy Principles for Information-Based Transparency. Government Information Quarterly, Volume 27, Issue 4, Pages 377-383

Stiglitz, J. E., Orszag, P. R., & Orszag, J. M. (2000). *The role of government in a digital age.* Commissioned by the computer \& communications industry association.

Suber, P. (2012). *Open Access.* MIT Press.

Tauberer, J. (2012). *Open Government Data: The Book.*

Ubaldi, B. (2013). *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives.* Tech. rep., OECD Publishing.

Umbrich, Jürgen, Sebastian Neumaier, and Axel Polleres (2015). *Towards assessing the quality evolution of Open Data portals*. ODQ2015: Open Data Quality: from Theory to Practice Workshop. Munich, Germany (March 2015).

Van de Sompel, H., Sanderson, R., Nelson, M. L., Balakireva, L. L., Shankar, H., & Ainsworth, S. (2010). *An HTTP-based versioning mechanism for linked data*. arXiv preprint arXiv:1003.3661.

Vickery, G. (2011). *Review of recent studies on PSI re-use and related market developments*. Information Economics, Paris.

Wand, Y, & Wang, R. (1996). *Anchoring data quality dimensions in ontological foundations*. Comm. ACM 39, 11.

Wang, R. & Strong, D. (1996). *Beyond accuracy: What data quality means to data consumers*. J. Manage. Inform. Syst. 12, 4.

Whitmore A. (2014) *Using open government data to predict war: A case study of data and systems challenges,* Government Information Quarterly, Volume 31, Issue 4, October 2014, Pages 622-630, ISSN 0740-624X.

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). *Innovation with open data: Essential elements of open data ecosystems*. Information Policy, 19(1), 17-33.

Zuiderwijk, Anneke, and Marijn Janssen (2015). "Participation and data quality in open data use: Open data infrastructures evaluated." Proceedings of the 15th European Conference on eGovernment 2015: ECEG 2015. Academic Conferences Limited, 2015.

WEB REFERENCES

Web ref. 1. http://opendefinition.org/ (last visited on November 5, 2015).
Web ref. 2. http://census.okfn.org/ , (last visited on November 5, 2015).
Web ref 3. http://goo.gl/QlUT1d (last visited on November 5, 2015).
Web ref 4. https://goo.gl/yq4ElP (last visited on November 5, 2015).
Web ref. 5. http://sunlightfoundation.com/blog/2010/06/23/elenas-inbox/ (last visit on November 5, 2015)
Web ref. 6. http://nexa.polito.it/lunch-9 (last visit on November 5, 2015)
Web ref. 7. http://okfnlabs.org/bad-data/ (last visit on November 5, 2015)
Web ref 8. http://www.dati.gov.it/dataset (last visit on November 5, 2015)
Web ref. 9. https://www.opengovawards.org/Awards_Booklet_Final.pdf, (last visit on November 5, 2015)
Web ref. 10. http://en.wikipedia.org/wiki/European_Social_Fund, (last visit on November 5, 2015)
Web ref. 11. http://www.dati.gov.it/ (last visit on November 5, 2015)
Web ref. 12. http://blog.okfn.org/2013/07/02/git-and-github-for-data/_(last visited on November 5, 2015)
Web ref 13. http://dat-data.com/ (last visited on November 5, 2015)
Web ref. 14. https://code.google.com/p/google-refine/ (last visited on November 5, 2015)
Web ref. 15. http://datacleaner.org/, (last visited on November 5, 2015)
Web ref. 16. http://checklists.opquast.com/en/opendata , (last visited on November 5, 2015)
Web ref. 17. http://odaf.org/papers/Open%20Data%20and%20Metadata%20Standards.pdf , (last visited on November 5, 2015)

APPENDIX A

1. Metrics defined

Characteristic	Metric	Variables	Formula	Scale	Normalization	Alternative in literature
Traceability	Track of creation	s: Source dc: Date of creation	$tc = 2s + dc$	[0, 3]	$tcn = \frac{tc}{3}$	-
	Track of updates	lu: List of updates du: Dates of updates	$tu = lu + du$	[0, 2]	$tun = \frac{tu}{2}$	-
Currentness	Percentage of current rows	ncr: Number of not current rows nr: Number of rows.	$pcr = \left(1 - \frac{ncr}{nr}\right) * 100$	[0%, 100%]	$pcrn = \frac{pcr}{100}$	Several authors gave different definitions of timeliness and currency (B. Heinrich et al., p 5, 2009). One of the most used (adopted by methodologies DQA, COLDQ, CDQ), is timeliness defined as: Timeliness = (max (0; 1-Currency/Volatility)) (Batini et al., 2009). Other references: (Heinrich, 2002) & (Ballou et al. 1998)
	Delay in publication	da: Date of information availability dp: Date of publication sd: Start date of the period of time referred by the dataset ed: End date of the period of time referred by the dataset.	$da = ed + 1$ $dp = 1 - \left(\frac{dp - da}{ed - sd}\right)$	(-∞, 1]	$dpn = dp$	-
Expiration	Delay after expiration	ed: Expiration date cd: Current date sd: Start date of the period of time referred by the dataset ed: End date of the period of time referred by the dataset.	$dae = 1 - \left(\frac{cd - ed}{ed - sd}\right)$	(-∞, +∞)	$if(dae \leq 0)$ $daen = 0$ $else\ if(\ dae \leq 1)$ $daen = rs$	-

					$else\ if\ (dae > 1)$ $daen = 1$	
Completeness	Percentage of complete cells	nr: Number of rows nc: Number of columns ic: Number of incomplete cells ncl: Number of cells	$ncl = nr * nc$ $pcc = \left(1 - \frac{ic}{ncl}\right) * 100$	[0%, 100%]	$pccn = \frac{pcc}{100}$	Completeness with the "open world" assumption (i.e., assumption that in the schema not all the real world entities are represented). (Batini & Scannapieco, 2006)
	Percentage of complete rows	nr: Number of rows nir: Number of incomplete rows	$pcpr = \left(1 - \frac{nir}{nr}\right) * 100$	[0%, 100%]	$pcprn = \frac{pcpr}{100}$	-
	Percentage of standardized columns	ns: Number of columns with associated standards nsc: Number of standardized columns	$psc = \left(\frac{ns}{nsc}\right) * 100$	[0%, 100%]	$pscn = \frac{psc}{100}$	-
Compliance	eGMS compliance	s: Source dc: Date of creation c: Category t: Title d: Description *(if applicable)* id: Identifier *(if applicable)* pb: Publisher *(if applicable)* cv: Coverage *(recommended only)* l: Language (*recommended only)*	$egmsc = s + dc + c + t + 0.2(d + id + pb + cv + l)$	[0 - 5]	$egmscn = \frac{egmsc}{5}$	Interpretability (metric used in the Data Warehouse Quality - DWQ methodology), defined as: "Number of tuples with interpretable data, documentation for key values" (Batini et al., 2009), and (Jeusfeld et al. 1998).
	Five stars Open Data		This metric does not require any formula; the value assigned depends on the level of the scheme in which the dataset is.	[0, 5]	$fsodn = \frac{fsod}{5}$	-
Understandability	Percentage of columns with metadata	ncm: Number of column with metadata nc: Number of columns	$pcm = \left(\frac{ncm}{nc}\right) * 100$	[0, 100]	$pcmn = \frac{pcm}{100}$	-

	Percentage of columns in comprehensible format	ncuf: Number of columns in understandable format nc: Number of columns	$pcuf = \left(\frac{ncuf}{nc}\right) * 100$	[0%, 100%]	$pcufn = \frac{pcuf}{100}$	-
Accuracy	Percentage of syntactically accurate cells	nce: Number of cells with errors ncl: Number of cells	$pac = \left(1 - \frac{nce}{ncl}\right) * 100$	[0%, 100%]	$pacn = \frac{pac}{100}$	Semantic accuracy, in which are considered not only the values not belonging to a certain domain but also all the values that don't represent the real world entity correctly, e.g. incoherent values, typos in names (Batini & Scannapieco, 2006), (Heinrich, 2002), (Kaiser et al., 2007).
	Accuracy in aggregation	e: Errors sum s: Scale oav: Own aggregation value dav: Dataset aggregation value	$e = \sum_{i=1}^{n} \lvert dav_i - oav_i \rvert$ $ea = 1 - \left(\frac{e}{s}\right)$	(-∞, 1]	$if(ea \leq 0)$ $ean = 0$ $else\ if(ea \leq 0.9)$ $ean = 0.25 * ea$ $else\ if(ea \leq 0.95)$ $ean = 0.5 * ea$ $else\ if(ean \leq 0.999)$ $ean = 0.75 * ea$ $if(ea > 0.999)$ $ean = ea$	The metric "derivation integrity" in the TIMQ framework calculates the same thing but in a broader way, it is defined as "Percentage of correct calculations of derived data according to the Integrity derivation formula or calculation definition" (Batini et al., 2009) and (English, 1999). -

APPENDIX B (ONLINE MATERIAL)

1. Explorative questionnaire

TABLE 1. QUESTIONS

Id	Question	Possible Answers
A	What type of application did you develop?	Open answer
B	Did you use any datasets disclosed by Italian providers?	Yes - No
C	Could you list the datasets that you used?	Open answer
D	How would you overall evaluate the quality of Open Data?	(1) Very low … (5) Very high
E	How much easy was to understand data?	(1) Very difficult … (5) Very easy
F	On average, how much time did you spend to understand your datasets?	Open answer
G	How much useful was to read metadata in order to better understand data?	(1) Not useful at all … (5) Very useful
H	How would you evaluate the completeness of the data you used for developing your application?	(1) Very low … (5) Very high
I	How much did you modify your original idea to being able to use the open data?	(1) Not changed at all … (5) Totally changed
J	Was the data format clear?	(1) Not clear at all … (5) Crystal clear
K	Did you have to modify the data format in order to use the data into your application?	Yes - No
L	Did you find errors on data?	Yes- No
M	Which problems did you find working with open data?	Open answer
N	Which aspects of data quality would you like to improve?	Open answer

2. Datasets details (municipality level)

TOPIC	CITY	DESCRIPTION	URL
Resident citizens	Torino	Resident citizens, 2011	http://www.comune.torino.it/aperto/dati/demografia/residenti-anno-2011.shtml
	Bologna	Resident citizens of 19-24 years old by place of residence	http://dati.comune.bologna.it/node/371
	Firenze	Resident citizens by age profile	http://opendata.comune.fi.it/statistica_territorio/dataset_0091.html
	Milano	Resident citizens by gender and place of residence, 1999-2011	http://dati.comune.milano.it/dato/item/29
	Roma	Resident citizens by place of residence and quinquennial age profile, 2011	http://dati.comune.roma.it/download/popolazione-e-societa/popolazione-iscritta-anagrafe-municipio-e-classi-di-eta-quinquennali
Marriages	Torino	Marriages by rite and marital status, 2011	http://www.comune.torino.it/aperto/dati/demografia/matrimoni-secondo-rito-e-stato-civile-anno-2011.shtml
	Milano	Marriages in Milano, 2003-2011	http://dati.comune.milano.it/dato/item/138
	Firenze	Marriages and divorces	http://opendata.comune.fi.it/statistica_territorio/dataset_0084.html

Business Activities	Torino	Business activities, 2011	http://www.comune.torino.it/aperto/dati/att_comm/negozi/attivit-commerciali-anno-2011.shtml
	Roma	Business Activities in town, 31-12-2012	http://dati.comune.roma.it/download/esercizi-commerciali/esercizi-commerciali-presenti-sul-territorio-comunale-31-12-2012
	Milano	Business activities of Medium and big distribution	http://dati.comune.milano.it/dato/item/50