An Exploratory Empirical Assessment of Italian Open Government Data Quality

SoftEng – Technical Report No. 2014/1

Title:

An Exploratory Empirical Assessment of Italian Open Government Data Quality

With an eye to enabling linked open data

Authors:

Antonio Vetro', Marco Torchiano, Camilo Minotas Orozco, Giuseppe Procaccianti, Raimondo Iemma, Federico Morando

Politecnico di Torino
Department of Control and Computer Engineering
Software Engineering group

SOftEng
http://softeng.polito.it

An Exploratory Empirical Assessment of Italian Open Government Data Quality

With an eye to enabling linked open data

Antonio Vetro' [1,2], Marco Torchiano [1], Camilo Minotas Orozco [1], Giuseppe Procaccianti [1,3], Raimondo Iemma [4], Federico Morando [4]

[1] Politecnico di Torino (Italy), [2] Technische Universität München (Germany), [3] VU University Amsterdam (Netherlands), [4] Nexa Center for Internet and Society , Torino (Italy)

ABSTRACT

Context

The diffusion of Linked Data and Open Data in recent years kept a very fast pace. However evidence from practitioners shows that disclosing data without proper quality control may jeopardize datasets reuse in terms of apps, linking, and other transformations.

Objective

Our goals are to understand practical problems experienced by open data users in using and integrating them and build a set of concrete metrics to assess the quality of disclosed data and better support the transition towards linked open data.

Method

We focus on Open Government Data (OGD), collecting problems experienced by developers and mapping them to a data quality model available in literature. Then we derived a set of metrics and applied them to evaluate a few samples of Italian OGD.

Result

We present empirical evidence concerning the common quality problems experienced by open data users when using and integrating datasets. The measurements effort showed a few acquired good practices and common weaknesses, and a set of discriminant factors among datasets.

Conclusion

The study represents the first empirical attempt to evaluate the quality of open datasets at an operational level. Our long-term goal is to support the transition towards Linked Open Government Data (LOGD) with a quality improvement process in the wake of the current practices in Software Quality.

Keywords: Open Government Data, Empirical assessment, Data quality, Linked data, Quality model

1. Introduction

'Open data' are defined by the Open Knowledge Foundation as "data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike."[1] In a nutshell, the underlying idea is that a wider and easier circulation of datasets otherwise unavailable to the general public could entail interesting (and even unexpected) forms of reuse, and in general improve transparency and understanding of phenomena (Aichholzer & Burkert, 2004). In general, compared to proprietary models, the circulation and reuse of the so-called 'digital commons' is characterized - from a legal point of view - by lower restrictions. This aspect is supposed to ultimately foster participation, creativity and innovation (Hofmokl, 2010).

Beyond legal openness, attempts to increase meaningfulness and reusability, also require to represent and manage data so that it can be easily queried, enriched and linked with other data by anyone. This can be achieved, for instance, by fulfilling principles such as the ones proposed by Berners-Lee (Berners-Lee, 2011), focusing on the transition from the Web of Documents to the Web of Data.

In recent years, the open data approach has been adopted by a growing number of Public Administrations. In fact, public bodies collate and manage tremendous amounts of information. Releasing them as open can provide considerable added value, meeting a demand coming from all kinds of actors, and increasing the transparency of institutions (Stiglitz, Orszag, & Orszag, 2000) (Ubaldi, 2013). Since 2009, when the first governmental Open Data portal (data.gov, in the U.S.) was launched, more than 300[2] public data catalogs were made available, and this trend is supposed to continue even more significantly.

In this work we posit that, in order to be usable and effective, open data needs a certain level of quality, both intrinsic and contextual, i.e., both in the way they are presented to the final user and/or to intermediaries and in relation to their potential usage goal(s). We believe that a data quality model for open data and a set of actionable metrics are necessary tools to achieve data quality improvement. In particular, the "open" nature of a data set magnifies the implications of its contextual quality: open data are supposed to stimulate serendipitous reuse and generate unexpected mixes and matches; though this is impossible, unless virtually anybody (with the required technical skills) is allowed to access the data, understand them, port them to other systems, etc. Moreover, this should be possible without relying on implicit knowledge only available within the organization that produced the data.

According to Heath and Bizer (Heath & Bizer, 2011), in November 2010, government data accounted for 43.12% of RFD triples in datasets registered in the Linking Open Data Cloud[3]. At that time, about 12% (25/203) of the data sets in the Linking Open Data Cloud group on the Datahub.io[4] were government data, while currently almost 20% (67/337) of the data sets in the same group are tagged as "government" data. Thus, on the one hand, open government data represent a relevant share of linked open data and this percentage could easily grow, considering that just a small share of government data is already published as linked data[5]. At the same time, open government data have a huge potential of improvement,

[1] http://okfn.org/opendata/ , (last visited on September 16, 2013).

[2] http://datacatalogs.org/dataset includes 337 data catalogs (last visited on August 28, 2013).

[3] http://lod-cloud.net/, (last visited on September 16, 2013).

[4] http://datahub.io/group/about/lodcloud including 337 datasets (last visited on September 16, 2013).

[5] In Italy, according to the Italian national portal, less than 4.8% (365/7618) of open government data can be considered as linked open data. See http://www.dati.gov.it/content/infografica (last visited on September 16, 2013). Moreover, a small number of public administrations is publishing the majority of these linked data.

in terms of quality. Again following Heath & Bizer (Heath & Bizer, 2011), in 2010, open government data represented just 4.46% of linked open data in terms of links. This is consistent with the impression of the authors of this paper, according to which the quality of linked data published by public sector bodies is frequently poor, in particular because of the fact that published data tend to be self-referential and do not adhere to standards. Moreover, this situation makes it complex also for third parties to link open government data amongst them and with third-party data.

Therefore, we posit that assessing (and contributing to increase) the quality of open government data in general, and the quality perceived by third-party developers in particular, may represent one of the preconditions for assessing and increasing the number and quality of linked open data in general. For this reason we focus on the empirical evaluation of open data, and specifically open government data.

In this context, we first conducted an exploratory investigation to understand the perception of data quality by developers that used and integrated open data during hackathons. The analysis of results allowed us to identify the most urgent problems in open data quality. Afterwards, using a data quality model already defined in the literature, we defined metrics for the quality aspects mostly related to those problems and we used them to evaluate the quality of open datasets released by Italian municipalities with an eye to identifying the key quality features enabling the production of linked open (government) data.

The most prominent contribution of this work is a set of metrics valid for measuring intrinsic quality of open data and pursue quality improvement of disclosed open data (and its translation in a linked formalism in a later stage). In addition, our exploratory analysis of Italian Open Government Data allows us to deliver a set of acquired good practices, weaknesses and guidelines to data providers.

The remainder of the paper is organized as follows: Section 2 introduces the problem of bad data quality and shows anecdotal evidence. Section 3 discusses previous attempts in modeling data quality, however we were not able to find any empirical assessment at operational level. Thus this is, up to our knowledge, the first concrete empirical assessment of open government data at raw data level.

Goal and study design (Section 4) follow, then we enunciate the problems found in the questionnaire (Section 5) and the metrics defined for related quality aspects (Section 6).

The results of the empirical assessment are presented in Section 7 and their interpretation is discussed in Section 8. We summarize limitations and contributions in Section 9 and 10, providing as well our roadmap and useful recommendations for future works.

2. Motivations

The main goal of disclosing data is getting something new from it, e.g. a new visualization that offers interesting insights, or a transformation linking and correlating different datasets for knowledge discovery. Low data quality can negatively affect the value of open data by making its transformation difficult or even impossible. The available literature reports several examples in this respect.

(Allison, 2010) reports problems of accuracy, aggregation and precision in open government data, e.g., 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 about poor aggregation and precision comes from a project of Sunlight Foundation called Fortune 535, in which the organization used the personal financial disclosure forms that any Congressman in USA 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.) and Congress changed these ranges several times, so that it was impossible to create consistent time series (e.g., to analyze which members of Congress accumulated richness during public service).

Other examples of bad data quality concern format: for instance Tauberer (Tauberer, 2012) reports that the two chambers of the U.S. Congress disclosed their data using different formats. As a consequence merging or comparing data is much more difficult. Consistency of names could also arise in this case (because of different IDs for Members of Congress). Sunlight's labs director Tom Lee reports data quality problems in a blog post entitled "How Not to Release Data"[6]. 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 we performed on the city of Torino (Italy) open datasets, we discovered problems regarding absence of metadata, not reported measuring system for geographical information and missing data[7].

Some of aforementioned problems might be avoided by adopting simple countermeasures, such as automatic data insertion. However a more disciplined and organized way of collecting and disclosing data is needed. Open data is potentially used by a large crowd and bad quality data can negatively impact their economic value and their effects on governments transparency and efficiency, as shown above.

The problem of data quality is even more urgent if we consider the trend of opened datasets in the recent years: more and more governments and organizations are opening their data[8], following the path set out by the Freedom of Information Act by US President Obama.

[6] http://sunlightfoundation.com/blog/2010/06/23/elenas-inbox/ (last visited on September 16, 2013)

[7] http://nexa.polito.it/lunch-9 (last visited on September 16, 2013)

[8] http://datacatalogs.org/dataset (last visited on September 16, 2013)

In Italy, for instance, the law (section II of Decree n. 179 of October 18[9]) currently provides that any data and documents produced by public offices must be open, unless some narrow exceptions apply. At the same time, the Agency for Digital Italy produced its *Guidelines for Semantic Interoperability through Linked Open Data*[10]. However, such praiseworthy initiatives risks to be frustrated if data are disclosed without guidelines and models that ensure a high level of quality, mainly because insufficient quality levels may reduce reuse opportunities. In particular, since to date public administrations do not typically natively produce linked data or data in RDF format (instead, it is likely that data is 'RDF-ized' at a later stage), the quality of any linked open data they publish cannot be higher than the quality of the underlying open government data, but no guidelines for the empirical assessment of such quality are available.

3. Previous work in modeling web data quality

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, 2011). A 5-stars open dataset should comply to 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 expressing an aspect of data quality, the schema proposed by Tim-Berners Lee covers only a specific quality aspect, i.e. the format or encoding used to publish the data. In fact, a dataset can reach the 5 stars level while showing at the same time poor quality (e.g., containing missing or low precision data). The approach adopted in this paper has a broader scope, aiming at building a comprehensive quality model for open data grounded on the practical problems confronted by final users.

For instance, a possible additional feature for the 5-stars schema, recognized by its author himself, is the addition of a sixth star related to the presence of metadata, possibly retrievable from a major catalog. In this respect, in 2004, the UK government established a standard, called e-GMS[11], to define which metadata fields are mandatory for e-Government resources. That standard identified, e.g., Creator, Date, Subject Category, and Title as compulsory fields.

The issue of open data quality became preeminent in 2007, when in Sebastopol, CA a meeting to develop a set of principles of open government data was held. The participants, 30 open data and World Wide Web experts (among which Lawrence Lessig and Aaron Swartz) dubbed themselves as "Open Government Working Group". During that meeting, the group

[9] http://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2012-12-18&atto.codiceRedazionale=12A13277 (last visited on September 16, 2013)

[10] http://www.digitpa.gov.it/sites/default/files/allegati_tec/CdC-SPC-GdL6-InteroperabilitaSemOpenData_v2.0_0.pdf (last visited on September 16, 2013)

[11] http://www.esd.org.uk/standards/egms/ (last visited on September 16, 2013)

produced a list of 8 principles for Open Government Data[12], according to which data must be: (1) Complete, (2) Primary (i.e., as collected at the source), (3) Timely, (4) Accessible, (5) Machine processable, (6) Non-Discriminatory (i.e., available to anyone, without registration), (7) Non-Proprietary (e.g., in terms of format), and (8) License-free (i.e., not subject to legal restrictions which are not mandated by statutes). Finally, "compliance must be reviewable": a "contact person must be designated" and an "administrative or judicial court must have the jurisdiction to review whether the agency has applied these principles appropriately".

These principles provide the foundations to develop an assessment process to evaluate open data quality.

A high-level analysis of Open Government Data (OGD) is provided by Ubaldi (2013). The author proposes a framework to perform empirical analysis of OGD initiatives, based upon questionnaires and interviews. The framework represents a first step towards a quantitative assessment of OGD quality, however it does not provide a set of well-defined metrics to evaluate quality aspects.

Our approach adopts similar techniques for data collection to those defined by Ubaldi (Ubaldi, 2013), but differs in terms of assessment and evaluation. In our work, we address the issue of OGD quality inspired by what has already been done in the context of Software Engineering, that is, we rely on Data Quality Models and on measurements and metrics for quality improvement.

Our starting point is the SQuaRE (Software Quality Requirements and Evaluation) model (ISO/IEC, 25010 International Standard: Systems and software engineering -Systems and software Quality Requirements and Evaluation - System and software quality models, 2010), also known as the 25000 Standard Series, by the ISO/IEC. More specifically, we make reference to ISO 25012, which "defines a general data quality model for data retained in a structured format within a computer system" (ISO/IEC, 25012 International Standard: Systems and software engineering - Software Product Quality Requirements and Evaluation (SQuaRE)-Data quality model, 2008). The 25012 model is organized as a set of 15 characteristics, that represent attributes having a positive impact on data quality. These characteristics can be considered from two different viewpoints:

- Inherent or *internal*: the capability of a dataset to satisfy user needs when used under specified conditions;
- System-dependent or *external:* the capability of a dataset to satisfy user needs when used under specified conditions and in a specific computer system.

We list in *TABLE 1*the 15 characteristics, according to the specific viewpoint; in some cases, a characteristic may be considered under both of them, but it applies to different aspects. For example, *confidentiality* may be inherent (e.g., through data encryption) but also system-dependent (the system may or may not be remotely accessible).

[12] The 8 principles of Open Government Data http://www.opengovdata.org/home/8principles (last visited on September 16, 2013)

TABLE 1. SQUARE QUALITY CHARACTERISTICS

Characteristic	Inherent	System-dependent
Accuracy	X	
Completeness	X	
Consistency	X	
Credibility	X	
Currentness	X	
Availability		X
Portability		X
Recoverability		X
Accessibility	X	X
Regulatory Compliance	X	X
Confidentiality	X	X
Efficiency	X	X
Precision	X	X
Traceability	X	X
Understandability	X	X

This model, however, is very general and is not usable "as-is" to evaluate open data quality. For example, it does not take into account user expectations. For this reason, we also considered a different quality model, developed by Calero et al. (Calero, Caro, & Piattini, 2008), named Portal Data Quality Model (PDQM).

PDQM is a model focused on the user perspective, aimed at evaluating data quality for Web portals. It was developed from the results of a systematic literature review and included a list of 41 quality attributes. Those attributes were subsequently classified in terms of relevance with respect to Web portal functionalities and user expectations. This was done to assign the appropriate attributes in order to assess data quality (DQ), taking into account the DQ expectation of data consumers by functionality. From this process, a final list of 34 DQ attributes was elicited.

The attributes were partitioned by authors in 4 different categories:

1. *Intrinsic*: attributes that evaluate DQ as an intrinsic property. E.g. Accuracy, Objectivity;
2. *Accessibility*: attributes that evaluate the level of security of the target system. E.g. Accessibility, Security;
3. *Contextual:* attributes that evaluate DQ with respect to the context of the task at hand. E.g. Completeness, Relevance, Timeliness;

4. *Representational*: attributes that evaluate DQ in terms of how they are presented to the user and their ease of comprehension and understanding. E.g. Interpretability, Consistent Representation, Easy of Understanding.

Finally, our last reference is a model presented by Moraga et al. (2009) that tries to combine both of the previous models in a more flexible, easy-to-use model for data quality evaluation. This model is called SQuaRE-aligned Portal Data Quality Model (SPDQM). The goal of this latter model was to reconcile the power and the accuracy of SQuaRE, as an international standard, with the specific issues of the Web Portal data quality assessment.

The development process of this model started from merging the 15 characteristics of SQuaRE with the 34 of PDQM. Moreover, a new systematic review was conducted, which resulted in the identification of 39 new quality attributes. A refinement process was carried out, to filter out those characteristics that were overlapping or irrelevant for the purpose, and finally a set of 42 characteristics was identified (30 PDQM characteristics, 5 new characteristics and 7 characteristics from ISO/IEC 25012).

An interesting feature of SPDQM is that it retains both the viewpoints of SQuaRE (inherent-system dependent) and the categories of PDQM (Intrinsic, Operational, Contextual, Representational). This inclusiveness allows defining more precisely each characteristic and making the assessment process more straightforward. The final structure of the SPDQM model is presented in *Figure 1*. We used this model as starting point for our empirical evaluation.

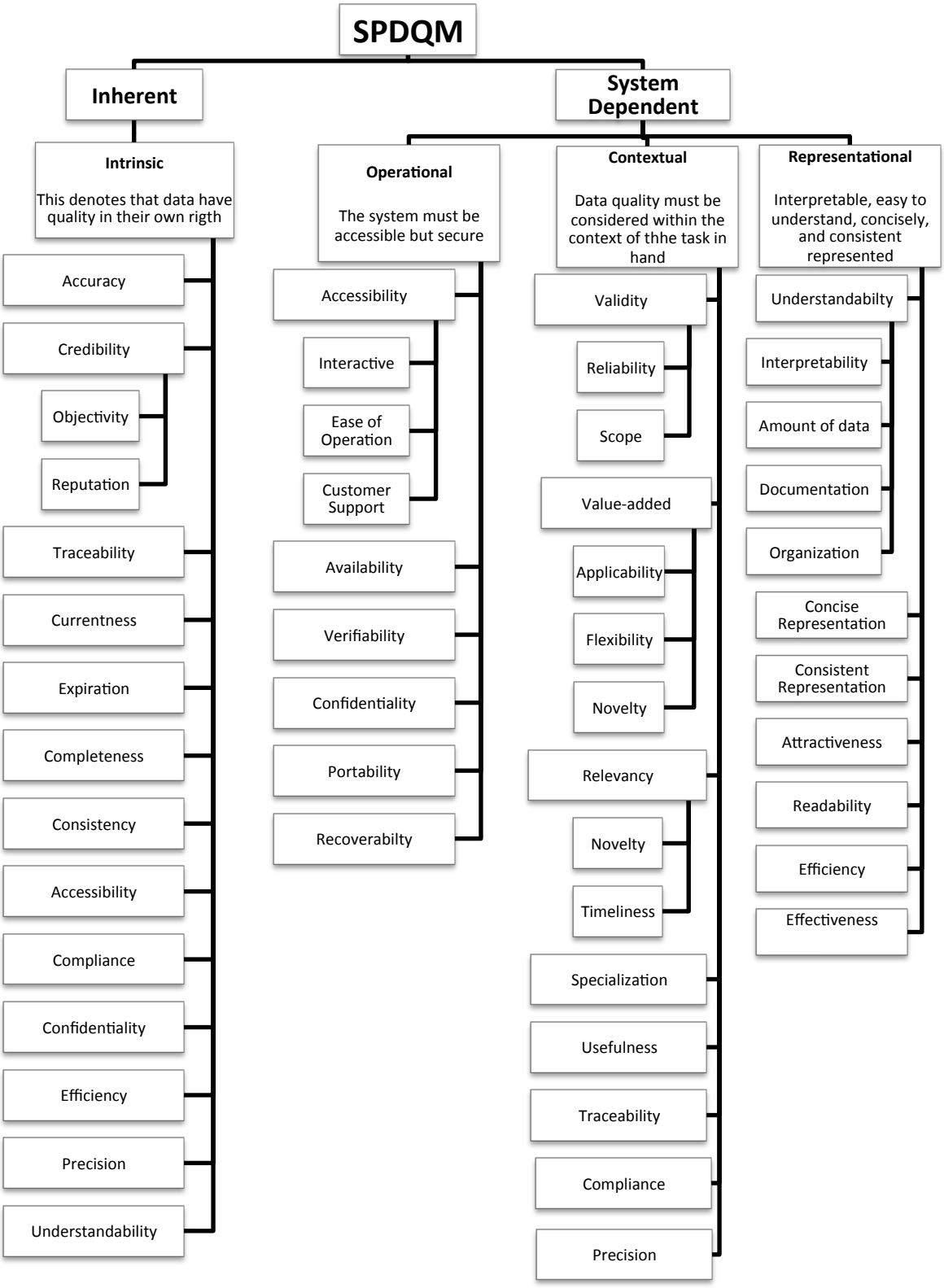

Figure 1. Structure of SPDQM

4. Study goals and design

We started this investigation having in mind two main immediate goals.

[G1] Firstly we aimed to understand which are the most recurring problems in open data quality, with an eye to identifying the key quality features enabling the production of linked open (government) data.

[G2] Secondly, starting from the most reported problems, we intended to define a set of metrics to measure the relevant related quality attributes. In fact metrics and measurement are conditions *"sine qua non"* for quality improvement. In practice, the first goal provides input for achieving the second goal.

Besides the two aforementioned goals, that are short terms goals achieved in this study, we enunciate also two medium term goals and a long-term goal, for which this study represents an initial step.

[G3] Build an improved model on open data quality grounded on the problems faced by final users in reusing, integrating and linking open data.

[G4] Take a snapshot of the current situation of Italian government data, packaged with a set of guidelines and best practices to embed quality checks in the processes of public administrations. Here the role of replication and automation is crucial.

[G5] In the long term we also aim at fine tuning the data quality model and the related guidelines and best practices, possibly coupling them with a partly automated process, with the aim of publishing "linkable" data. By "linkable" we mean RDF data (or, in any case, data represented by stable URIs), which conforms to standards, where these exist, or are at least documented with exhaustive human readable metadata –and machine processable--, so that third parties can link such data to the rest of the Web of Data.

The empirical methodology used to achieve the first goal (G1) is that of exploratory surveys. In particular, we conducted three surveys in order to understand the perception of data quality by final users' perspective. The first survey was used as a pilot that helped designing the other two. The result of the surveys consists in a list of common issues encountered by developers when dealing with open (government) data (see Section 5).

Goal G2 is achieved on the basis of the outcome of the surveys. In particular we performed a mapping of the most relevant issues identified (G1) onto the data quality characteristics defined by the SPDQM. Once the practically relevant characteristics were identified, a set of metrics was developed for assessing such characteristics. The definition of metrics -- an original contribution from this paper -- took into account both the abstract interpretation of the characteristics and the concrete issues sampled with the survey (see Section 6).

The initial application of the metrics on a selected group of open government data (see Section 7) represents the first step in achieving goal G4.

As far as goal G3 and G5 are concerned, we plan to improve and replicate the survey on a larger and more structured group of developers, for example in Italy with the community "spaghettiopendata"[13], which develops applications using Italian GOD and counts more than 600 subscriptions. War stories, focus groups and structured interviews will follow the survey to deepen specific aspects related to Linked Open Data.

The overall process and the relationship of the activities with the above goals are depicted in Figure 2.

[13] http://www.spaghettiopendata.org/ (last visited 24 September 2013)

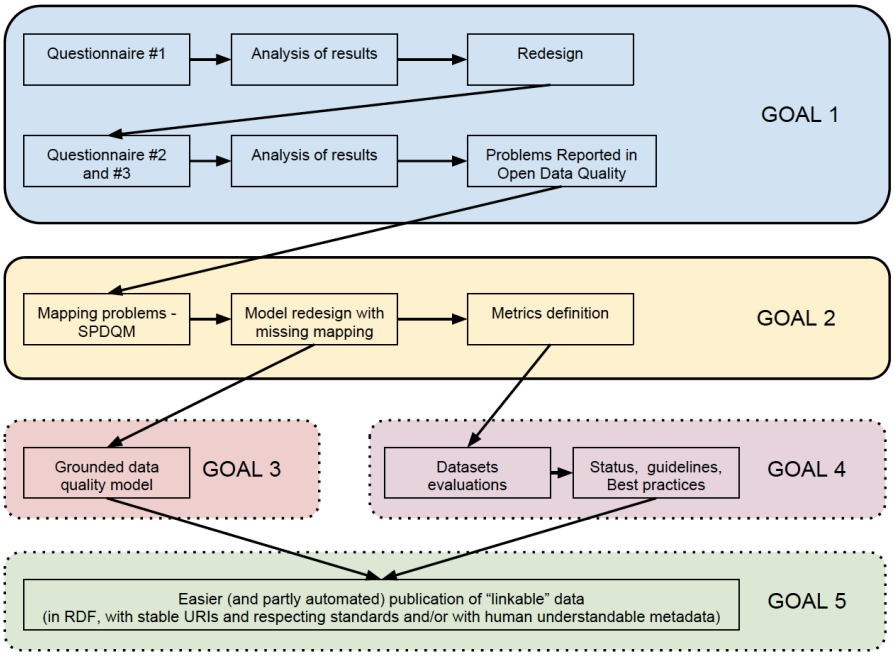

Figure 2. Short and long term goals of the study

The first activity carried on was a survey (Questionnaire #1 in Figure 2) administered to the participants in a seminar[14] on Open Data Quality held by the first two authors of the paper at hand. The participants were developers, representatives of regional institutions, and open data experts. The questionnaire consisted of six items: the first four required an answer on a five points ordinal scale and served the purpose of assessing the experience of the respondent, the fifth focused on the perception of quality of used open datasets on an ordinal scale, the last one consisted in an open ended question concerning the relevant aspects of open data quality. The questionnaire was in Italian language, questions and possible answers are translated and reported in Appendix A).

The outcomes of the first survey were used to design a new version of the instrument that was used in the two following investigations (Questionnaire #2 and #3 in Figure 2).

In particular we aimed at a more homogenous set of participants, namely developers, and therefore we defined a set of items addressing more specific aspects of data quality.

The questionnaire instrument, reported in *TABLE 2*, was administered to the participants of two different hackatons on Open Data:

1. Android University Hackaton 2013[15], in 8 different Italian Universities;
2. Hackathon Open Data[16] - Nexa Center for Internet & Society, in Torino and Alessandria.

[14] http://nexa.polito.it/lunch-9 , held at Nexa Center for Internet & Society located at Politecnico di Torino, Italy. (last visited on September 24, 2013)

[15] https://googledrive.com/host/0B_Ti0S8vWiPjMjZEMldJRnBXV1k/ (last visited on September 16, 2013)

[16] http://nexa.polito.it/lunch-13 (last visited on September 24, 2013)

TABLE 2. QUESTIONNAIRE #2 AND #3

Id	Question	Possible Answers
Q2.A	What type of application did you develop?	Open answer
Q2.B	Did you use any datasets disclosed by Italian providers?	Yes - No
Q2.C	Could you list the datasets that you used?	Open answer
Q2.D	How would you overall evaluate the quality of Open Data?	(1) Very low … (5) Very high
Q2.E	How much easy was to understand data?	(1) Very difficult … (5) Very easy
Q2.F	On average, how much time did you spend to understand your datasets?	Open answer
Q2.G	How much useful was to read metadata in order to better understand data?	(1) Not useful at all … (5) Very useful
Q2.H	How would you evaluate the completeness of the data you used for developing your application?	(1) Very low … (5) Very high
Q2.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
Q2.J	Was the data format clear?	(1) Not clear at all … (5) Crystal clear
Q2.K	Did you have to modify the data format in order to use the data into your application?	Yes - No
Q2.L	Did you find errors on data?	Yes- No
Q2.M	Which problems did you find working with open data?	Open answer
Q2.N	Which aspects of data quality would you like to improve?	Open answer

In next section we provide a quantitative summary of the closed-answer items of the questionnaire (i.e. those with Likert scales, yes/no, and time). Moreover we perform a coding of answers to open questions (Q2.M and Q2.N) to extract useful insights concerning data quality perception, which were useful in providing the basis to define the second part of the study.

5. Typical issues experienced by open data users

At the time of completing this paper, we collected 6 answers from the first survey and 9 answers from the second one.

Given the exploratory and qualitative nature of the questionnaire, the low number of answers represents a limitation in generalizing its results, but not in terms of internal validity. We further discuss this point in Section 9.

TABLE 3 reports the answer frequencies for each of the levels of Likert scale and graph, *TABLE 4* contains descriptive statistics for question Q2.F, while *TABLE 5* reports answers for questions yes/no (Q2.K and Q2.L).

TABLE 3. FREQUENCIES OF ANSWERS TO QUESTIONS ON LIKERT SCALE

Id	Question	Mean	Frequencies					
			1	2	3	4	5	
Q2.D	How would you overall evaluate the quality of Open Data?	2.33	3	1	4	1	0	
Q2.E	How much easy was to evaluate data?	3.11	2	2	1	4	1	
Q2.G	How much useful was to read metadata in order to better understand data?	3.00	2	2	3	1	2	
Q2.H	How would you evaluate the completeness of the data you used for developing your application?	2.44	3	2	2	3	0	
Q2.I	How much did you modify your original idea to being able to use the open data?	3.11	2	3	1	1	3	
Q2.J	Was the data format clear?	3.22	1	2	3	2	2	

TABLE 4. ANSWERS TO QUESTION Q2.F

Id	Question	Mean	Min	Max
Q2.F	On average, how much time did you spend to understand your datasets?	1h51m	10m	4h

TABLE 5. ANSWER TO QUESTIONS Q2.K AND Q2.L

Id	Question	Yes	No
Q2.K	Did you have to modify the data format in order to use the data into your application?	78%	22%
Q2.L	Did you find errors on data ?	56%	44%

The perceived quality of open data is generally low (average 2.33), and we observe problems about data completeness (Q2.H, average 2.44). Data format is in the middle of the Likert scale (average 3.22), however 78% of respondents had to modify it to being able to use the data. Metadata were not so useful (Q2.G, average 3), and since Q2.E shows that understandability is medium (3.11), probably this is because metadata are not providing useful guidance. Moreover, the average time spent to understand data (almost 2h) further support this fact. In addition, answers to question Q2.L highlight accuracy problems: in more than half of the cases errors were found in datasets. Coding of answers to question Q2.M ("Which problems did you find working with open data?") reported the following problems and frequencies:

- Incomplete data (5)
- Format (4)
- Traceability (3)
- Incongruence (2)
- Non uniformity (1)
- Update (1)
- Interface (1)

Coding of answers to question Q2.N ("Which aspects of data quality would you like to improve?") reported the following aspects and frequencies:

- Format (6)
- Completeness (5)
- Traceability (2)
- Congruence (1)
- Update (1)
- Metadata (1)

Answers Q2.N correspond to the problems reported, with the exception of interface issues, which therefore will not be considered at this step.

6. Metrics for quantitative evaluation

TABLE 6 summarizes the type of problems emerging from the surveys and links them to the data quality of 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.

Part of the mapping was straightforward: the codes "Incomplete data", "Traceability", "Congruence", "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 code "Update" could refers both to time validity and data obsolescence, for this reason it refers to both expiration and currentness.

Some discussion has to be done for code "Metadata". Answers to the questionnaire showed that developers encountered understandability problems, and our theory is that one of the reasons is that metadata do not provide useful guidance (metadata was not useful according to the questionnaire). Although we could not test this cause relationship, we believe that is safe and reasonable to map the code "Metadata" with Understandability. In addition, it is also mapped to compliance due to the existence of a standard for metadata in open government data sets (see next section).

Finally, code "Format" had no clearly corresponding quality characteristic. We mapped it to compliance and we measured it with the compliance to the 5 Stars Open data format scheme from Tim Berners-Lee (Berners-Lee, 2011).

Afterwards, for each of the selected quality attribute we defined metrics to measure the related quality aspect at dataset level. We took into account both the definition of quality characteristic and, whenever possible, the type of problem reported by developers. TABLE 7 contains the metrics we defined for each of the selected quality attributes. We report name and descriptions, while the formulas used to compute them are sown in Appendix A).

TABLE 6. PROBLEMS FOUND IN THE SURVEY AND LINKS TO SPDQM

Problem found in survey	**Related quality characteristic**	**Intrinsic**	**System-dependent**
Incomplete data	Completeness	X	
Format	Compliance		
Traceability	Traceability	X	X
Congruence	Consistency	X	

Update	Expiration, Currentness	X	
Metadata	Compliance, Understandability		
Errors	Accuracy	X	
High time to understand data	Understandability	X	X

TABLE 7. METRICS DEFINED AND DESCRIPTION

Characteristic	Metric	ID	Description
Traceability	Track of creation	TC	Indicates the presence or absence of metadata associated with the process of creation of a dataset.
	Track of updates	TU	Indicates the existence or absence of metadata associated with the updates done to a dataset.
Currentness	Percentage of current rows	PCR	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	DP	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	Date of expiration defined	DED	Indicates the existence or absence of metadata related to the date of expiration of a dataset.
	Delay after expiration	DAE	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	PCC	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	PCPR	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	PSC	Indicates the percentage of standardized columns in a dataset. It just considers the columns that represents some kind of information that has standards associated with it (i.e Geographic information).
	EGMS-Compliance	EGMSC	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	FSOD	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	PCM	Indicates the percentage of columns in a dataset that have associated descriptive metadata. This metadata is important because it allows to easily understand the information of the dataset and the way in which it is represented.
	Percentage of columns in understandable format	PCUF	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	PAC	Indicates the percentage cells in a dataset that have correct values according to the domain and the type of information of the dataset.
	Error in aggregation	EA	Indicates the ratio between the error in aggregation and the scale of data representation. This metric only apply 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.

7. Quantitative assessment of datasets quality

A. Datasets analyzed

In order to allow comparisons, we searched through Italian open government portals for datasets present in more than one portal . We selected the municipality level for two reasons: 1) it permitted us avoid aggregation problems at higher institutional levels; 2) being this the most fine grained level, we wanted to maximize the possibility to find datasets on common topics.

We ended up our search with dataset about three topics: resident citizens, marriages, and business activities. *TABLE 8* shows which dataset type was found in which city, while details and URLs for each datasets are reported in Appendix A).

TABLE 8. DATASETS TOPICS AND CITIES

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

B. Evaluation of Results

We report in Figure 3 the results of the datasets evaluation using the metrics defined in Section 6. The measures have been normalized to the interval [0,1] in order to allow comparison between the different datasets. In certain cases (e.g., DED, DAE and PCM) the metrics were not applicable or undefined. A metric is considered not available (NA) if it measures a characteristic that makes no sense applied to the dataset, for example, the error of aggregation (EA) is not applicable if there is no column that aggregates value from other columns. While, a metric is considered undefined (ND) when there is not enough information to compute it: for example, the delay in publication is undefined when the publication date is missing either on the web site or within the metadata. ND data can be considered equivalent to 0, so it will be in correspondence to a missing column in Figure 3, while NA is represented with an X in the corresponding missing column.

Given the exploratory nature of this work and the limited number of datasets manually verified, we don't perform a strict statistical analysis with hypothesis tests as a regular empirical analysis would require. In fact we aim to understand which are the strong and weak quality aspects in the different datasets, in provision for Goal 4 (see Figure 2 in Section 4). For this reason we organize results in three categories:

- **Acquired good practices**, i.e. metrics that are generally high in all datasets
- **Quality aspects to be improved**, i.e. in relation to metrics that are generally low in all datasets
- **Discriminant factors**, i.e. metrics that change significantly with respect to the data source and determine on which quality aspects a dataset is different from the others

The three categories and the corresponding metrics are reported in *TABLE 9*. FSOD metric (Compliance) is not reported in any group because all datasets are level 3 of Tim Berners Lee scale, exactly in the middle. We provide an interpretation of results and we articulate a deep discussion in Section 8.

TABLE 9. RESULTS OVERVIEW

CATEGORY	METRIC	QUALITY ATTRIBUTE
Acquired good practices	Track of creation (tc)	Traceability
	Percentage of current rows (pcr)	Currentness
	Percentage of complete cells (pcc)	Completeness
	E-GMS compliance (egmsc)	Compliance
	Error in aggregation (ea)	Accuracy
	Percentage of standardized columns (psc)	Compliance
Quality aspects to be improved	Track of updates (tu)	Traceability
	Percentage of columns with metadata (pcm)	Understandability
Discriminant factors	Delay in publication (dp)	Currentness
	Date of expiration defined (ded)	Expiration
	Percentage of complete rows (pcprn)	Completeness
	Percentage of columns in understandable format (pcuf)	Understandability
	Percentage of accurate cells (pac)	Accuracy
	Delay after expiration (dae)	Expiration

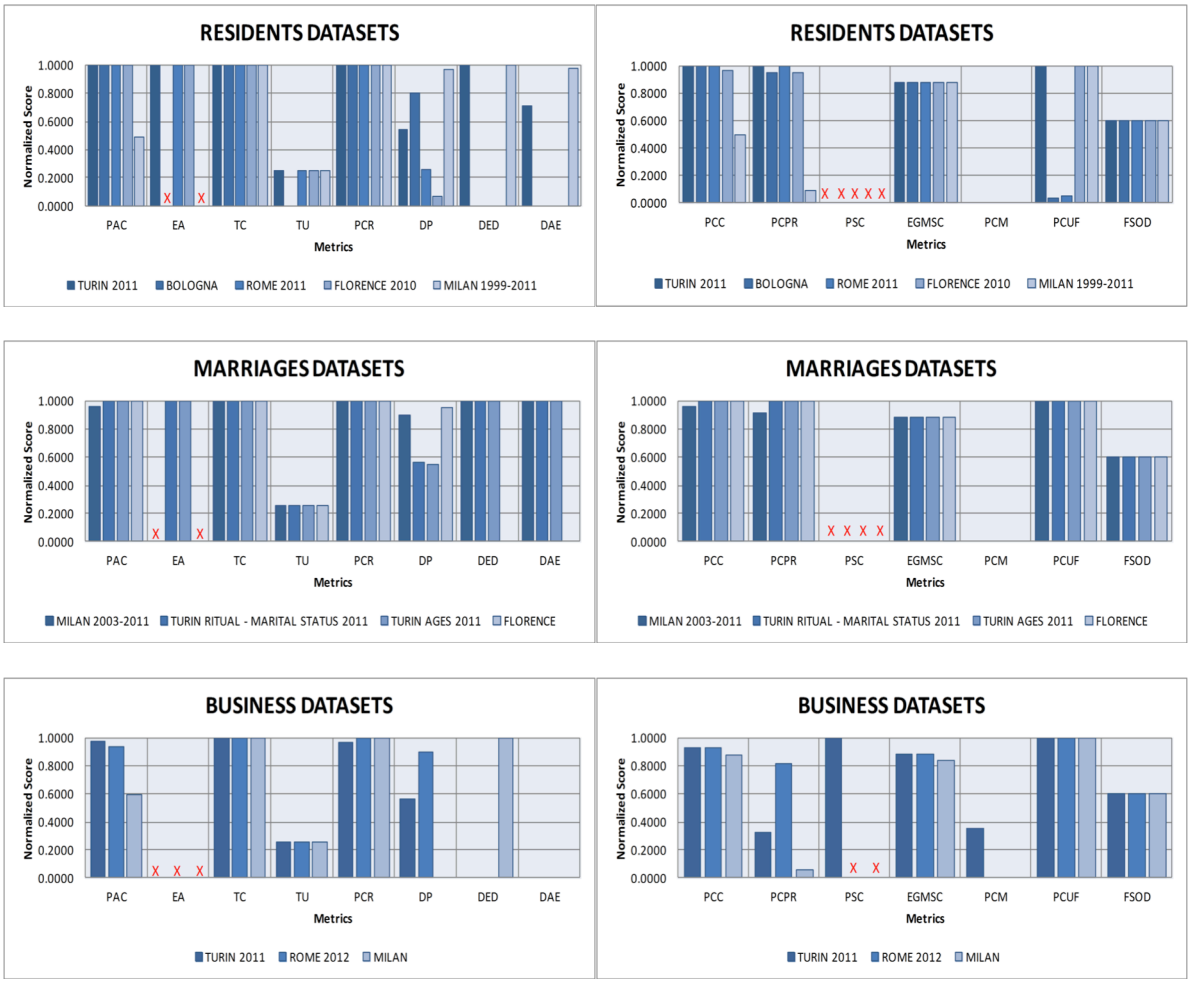

Figure 3. Results of datasets quality assessment

8. Discussion

We organize the discussion in two subsections: general observations (A) and specific comments concerning the path toward a broader publication of Linked Open Government Data (B).

A. General observations

In the previous section we categorized the metrics that consistently obeyed to a given pattern in acquired good practices and common weaknesses. On the contrary, metrics with no common pattern can be considered as discriminant factors that differentiate datasets by specific quality attributes. Since some quality aspects have more than one metric associated, we have situations in which the same quality characteristic has both an acquired good practice and a common weakness.

For example, regarding traceability, the dataset creation information is always present, however there is no information available on data modifications and updates. This is also partly reflected by one of the problems emerged from developers answers ("Update" problem category) and the discussion emerged from the seminar where we presented the pilot questionnaire. A possible solution might be to apply 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 grey literature. For instance, Rufus Pollock, founder and co-Director of the Open Knowledge Foundation (OKF), addresses this problem in a blog post[17] 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. This solution, already applied in some OKF projects, improves data traceability and encourages collaboration. The author also suggests some additional features like adding data format description through JSON[18] and basic scripts to enhance data accessibility and understandability.

Another quality aspect related to "Update" is data currentness: our assessment indicates that data is generally current, although it is not always released as soon as it is available. And, as a matter of fact, expiration is classified as discriminant factor. This aspect might be related to the lack of disclosure planning provided by the Italian institutions.

Regarding the completeness, the quantity of complete cells is high, however certain datasets seem to concentrate empty/non valid cells in specific rows (% of complete rows, PCRP metric). The percentage of columns in a understandable format is also a discriminant factor, which makes certain datasets more comprehensible than others. In addition, although the use of standardized metadata fields is an acquired good practice (when a standard is present, up to our knowledge), not all columns have metadata associated, which also affects understandability. A possible explanation: since existing standards do not cover all type of data, there is a gap that is not filled with any additional metadata description.

We did not find any aggregation issue (when assessment was possible), however data is not free from errors (as it was also remarked by developers), which means that values are sometimes inconsistent with their domain. For instance, the business activities dataset in

[17] http://blog.okfn.org/2013/07/02/git-and-github-for-data/ (last visited on September 16, 2013)

[18] http://data.okfn.org/standards/data-package (last visited on September 16, 2013)

Torino has the following domain[19] for column COD_COMP: {CFE, CF, E, EP}. However the column contains the following values: {AE, CF, EP}; notably the code AE is not present in the metadata schema. In this specific case, but also in a more general perspective, a feedback mechanism to take note of the errors found by the users might be useful, perhaps in conjunction with a 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 had the rights to modify data? How is the process of re-validation managed? How much resources are required to supervise the users' feedback mechanism? These open questions introduce yet another consideration: the necessity of an infrastructure to better handle the process of opening data management. Recent research works (Abecker, Heidmann, Hofmann, & Kazakos, 2013)) and existing solutions (e.g., Data.gov Dataset Management System[20]) are examples of infrastructures for managing data collection, selection, harmonization, transformation and export. Although our vision is much broader and consists of having a middleware to handle and assist the data provider not only on the above mentioned steps but also in performing quality checks. Metrics and measurements are enablers for setting up automatic quality gates, a practice that is largely adopted by software process improvements initiatives (see, for instance, CMMI[21]). Similarly to what happens in pre-release VV activities in software development (analysis of code, reviews, testing, etc.), quality checks might theoretically prevent disclosure of low quality data.

B. The path towards Linked Open Government Data

The graph on *Figure 4* shows the distribution of Italian datasets according to the 5-stars open data benchmark, during the period from March 2012 to September 2013. Less than 5% of open datasets is in the 5 stars level, about 25% is at 4 stars level, but the majority of disclosed data (63%) only reaches 3 stars level[22]. In the paper at hand all datasets are classified as 3 stars. These figures seem to clearly indicate that there is still room for integration between dataset, and the path towards Linked Government Open Data is still to be properly paved.

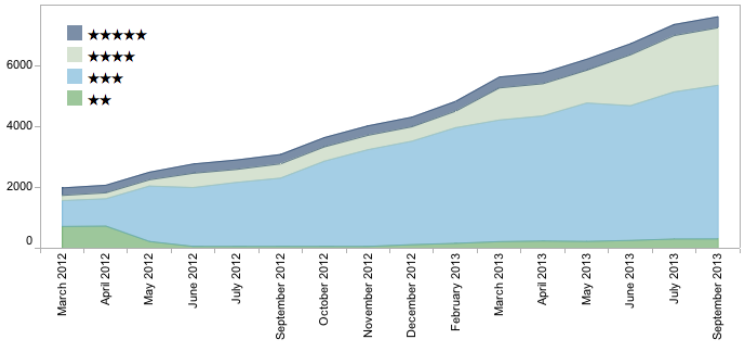

Figure 4. Distribution of Italian datasets according to 5 stars open data[23]

[19] http://www.comune.torino.it/aperto/bm~doc/attivita_commerciali_2011.pdf (last visited on September 16, 2013)

[20] https://dms.data.gov/ (last visited on September 16, 2013)

[21] http://cmmiinstitute.com (last visited on September 16, 2013)

[22] Notice that data described as "4 stars" in the table do not always fully respect the RDF formalism: for instance, the data available at http://www.dati.piemonte.it/rdf.html are properly serialized in RDF/XML, but their URIs are not dereferenceable (they return HTTP 404 error). Therefore, these data are not easily "linkable" in the sense described in the paper at hand.

[23] http://www.dati.gov.it/content/infografica#Quanti sono i dati aperti in Italia? , last accessed: 23 September 2013

Two basic obstacles in this regard are the lack of semantic orchestration between Open Government Data initiatives and the consequent lack of common semantic even within analogous and akin PA (as a matter of fact, see the example of the two Congress Chambers mentioned in Section 2). We are aware of ongoing projects, such as the Italian[24] Italia.gov.it and the European Open-DAI[25] (Opening Data Architectures and Infrastructures of European Public Administration), which aim at building a systematic view of the different government open data initiatives and setting up integration mechanisms through service oriented architectures. More generally, several relevant initiatives at the European level are linked to the ISA Programme[26], e.g., SEMIC - Semantic Interoperability Community. In addition, the Open Government Initiative aims at creating a model language for municipal governments. However region or nation-wide orchestration and semantic schemas are still work in progress and perhaps not yet a realistic path for many countries, for organizational and economical reasons. In the meanwhile, we believe that empirically extracting semantic schemas from datasets on common topics is a reasonable milestone in the path toward integrations of open datasets and the migration to Linked Open Government Data (LOGD), which is still far from being common practice. As a matter of fact, so far LOGD is only adopted by Data.gov.uk and it is just a recommendation –i.e., with no enforcement-- of the Australia Government Open Access and Licensing Framework[27].

Proper format and semantics are necessary enablers to LOGD, however they are not sufficient conditions. Our empirical assessment, although limited in scope, showed several other problems, especially in understandability, accuracy and completeness, which severely impair data integration. For instance, the percentage of columns with associated metadata is a poor quality aspect transversal to many datasets. This is especially critical, since government data are published under tight budget constraints and it would be frequently desirable to publish data following the RDF formalism, leaving the third parties at least part of the challenge related with their linking to the rest of the Web of data. However, this is not easily doable, unless the published data score very high in terms of understandability (possibly leaving to third parties the formal expression of a semantic which is clearly spelled at least for human beings). On the other hand, as we have already underlined in precedent discussion, when metadata standardization is available, it is used. This fact might still be a consequence of the absence of a reference semantic schema.

An obstacle of different nature is represented by accuracy and completeness: incomplete cells can lead to missing RDF triples, wrong accuracy translates not only into wrong content but also into wrong URIs, making in practice the 5th star useless. Finally, data currentness and expiration problems can be easily mitigated by an effective use of "Expires" header, while improved traceability is achievable with a more comprehensive use of metadata fields.

9. Limitations

This study is a first and partial attempt towards objective, reproducible and scientifically-based quality assessment of disclosed government data. Because of the current state of art of the open government datasets, we are one step before the realization of linked data: therefore the exploratory nature of the study impacts generalizability and reliability of findings. Some

[24] http://www.italia.gov.it/il-progetto (last visited on September 16, 2013)
[25] http://www.open-dai.eu/ (last visited on September 16, 2013)
[26] ISA – Interoperability Solutions for European Public Administrations: http://ec.europa.eu/isa/ (last visited on September 16, 2013)
[27] http://www.ausgoal.gov.au/ausgoal-qualities-of-open-data

of the threats, which affect the validity, are listed below according to (Wohlin et al., 2012) together with mitigation strategies and comments.

- Lack of generalizability. Due to time and effort required by manual evaluation, but also due to the difficulty of finding comparable datasets even in a large repository as the Italian OGD portal, the number of datasets evaluated is small and cannot represent the current health of the Italian OGD, which is our future goal (see Section 4). However the evaluation process is easily reproducible, but future work is necessary to make it scalable, introducing automatic measurements.
- Lack of metrics validation. The selected metrics might not properly represent the quality aspects, introducing construct threats to the validity of the study. Nevertheless, the assessment confirmed some of the problems reported by the developers, such as understanding and incompleteness. A rigorous validation would require using the metrics to evaluate the same datasets used by those developers who filled the questionnaire. Although this is a necessary step for the sake of validation, for practical reason it couldn't be performed in this first exploration.
- No statistical significance in results. We are aware of this conclusion threat, however having such kind of results is not yet a goal of this research, given the very initial status and explorative nature.

10. Conclusions and future work

We conducted an exploratory analysis of open data with two goals: i) understanding which are the practical problems experienced by open data users and ii) building a set of metrics to identify them and get an objective and reproducible quality assessment, also addressing some preconditions of the evolution into LOD.

We focused on Open Government Data and collected problems experienced by developers who participated in two hackathons. Afterwards, we mapped the quality problems to the quality characteristics described by the SPDQM quality model and derived a set of metrics as a first step towards the automatic and reproducible identification of issues.

We applied the defined metrics to evaluate a small set of Italian Open Government Data and extract common patterns and variations. We packaged them as contribution to state of the art in acquired good practices, common weaknesses, and discriminant factors among the datasets evaluated.

The study represents, up to our knowledge, the first attempt to empirically evaluate the quality of open datasets at operational level. Our medium term goal is to enlarge SPDQM, empirically building a more comprehensive quality model, better tailored to linked open data. Replications of this study on a larger set of data will be necessary, however some previous steps are necessary: the validation of the metrics and the automation of their computation are the most urgent.

In the longer run, this study paves the path toward a quantitative assessment of open data quality and will be the basics to properly guide the switch towards the use of LOD, with particular attention to the Government domain. Using the metrics proposed we were able to discover quality problems that constitute practical obstacles to properly get open government data in RDF and to link it to other resources. However, more effort is necessary to elaborate, validate and automate the solution proposed.

The problems reported by the developers and the implications of our vision produce input and ideas for future research work, such as the necessity of a middleware to better handle the disclosure process by means of quality gates, the inclusion of specific criteria for data integration and LOD, or even guidelines for assessing the quality of web services and APIs, including the quality of the related documentation.

Aknowledgments

We would like to sincerely thank the participants to the surveys for their valuable feedback. This study wouldn't have been possible without their collaboration.

REFERENCES

Abecker, A., Heidmann, C., Hofmann, C., & Kazakos, W. (2013). Towards a Middleware for Data Management in Support of Open Government Data. *Proceedings of the 27th Conference on Environmental Informatics - Informatics for Environmental Protection, Sustainable Development and Risk Management. 1*, pp. 674-681. Shaker Verlag GmbH.

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

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.

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

Berners-Lee, T. (2011). *Linked data-design issues (2006).* Tech. rep., W3C.

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

Ericson, J. (2010). Net expectations-what a web data service economy implies for business. *Information Management , 20* (1), 16.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology , 1* (1), 1-136.

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

ISO/IEC. (2010). *25010 International Standard: Systems and software engineering - Systems and software Quality Requirements and Evaluation - System and software quality models.* Tech. rep., ISO/IEC.

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

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).

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 by Joshua Tauberer.*

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

Wohlin, C., Runeson, P., Hst, M., Ohlsson, M. C., Regnell, B., & Wessln, A. (2012). *Experimentation in software engineering.* Springer Publishing Company, Incorporated.

APPENDIX

A) Pilot questionnaire

Id	Question	Possible Answers
Q1.A	How many years of experience do you have on open data ?	0, 1, 2, 3, 4, >4
Q1.B	How many applications did you develop using open data ?	0, 1, 2, 3, 4, >4
Q1.C	If you developed at least one application, how many different data sets did you integrate ?	1 (no integration), 2, 3, 4, >4
Q1.D	If you developed at least one application, from how many open data providers did you take data ?	1, 2, 3, 4, >4
Q1.E	How would you overall evaluate the quality of Open Data ?	(1) Very low (2) Low (3) Adequate (4) High (5) Very high
Q1.F	Which are, in your opinion, the most important aspects of open data quality ?	Open answer, up to 5 aspects

B) Metrics defined

Characteristic	Metric	Variables	Formula	Scale	Formula normalization
Traceability	Track of creation(tc)	s: source dc: date of creation	$tc = 2s + dc$	[0, 3]	$tcn = \frac{tc}{3}$
	Track of updates(tu)	lu: List of updates du: dates of updates	$tu = lu + du$	[0, 2]	$tun = \frac{tu}{2}$
Currentness	Percentage of current rows(pcr)	ncr: Number of not current rows nr: Number of rows.	$pcr = \left(1 - \frac{ncr}{nr}\right) * 100$	[0%, 100%]	$pcrn = \frac{pcr}{100}$
	Delay in publication(dp)	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	Date of expiration		If the metadata exists the value of the metric is 1, otherwise it is	[0, 1]	$dedn = ded$

	defined(ded)	None	0.		
	Delay after expiration(dae)	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(pcc)	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%]	$pccn = \frac{pcc}{100}$
	Percentage of complete rows(pcpr)	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(psc)	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(egmsc)	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]	$egmscn = \frac{egmsc}{5}$
	Five stars open data		This metric does not require any formula, the value assigned depends on the level of the	[0, 5]	$fsodn = \frac{fsod}{5}$

		scheme in which the dataset is.			
Understandabity	Percentage of columns with metadata(pcm)	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 understandable format(pcuf)	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 accurate cells(pac)	nce: Number of cells with errors ncl: Number of cells	$pac = \left(1 - \frac{nce}{ncl}\right) * 100$	[0%, 100%]	$pacn = \frac{pac}{100}$
	Error in aggregation(ea)	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]	$if(ea \leq 0)$ $ean = 0$ $else\ if(ea \leq 0.9)$ $ean = 0.25 * ea$ $else\ if(ea \leq 0.95)$

$$ean = 0.5 * ea$$

$$else\ if (ean \leq 0.999)$$

$$ean = 0.75 * ea$$

$$if (ea > 0.999)$$

$$ean = ea$$

C) Datasets details

TOPIC	CITY	DESCRIPTION	URL
Resident citizens	Torino	Resident citizens, 2011	http://www.comune.torino.it/aperto/dati/demografia/residenti-anno-2011.shtml
	Torino	Resident citizens by Age and birthplace, 2011	http://www.comune.torino.it/aperto/dati/demografia/et-e-regione-di-nascita-dei-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
	Roma	Resident citizens by gender and age, 2006-2011	http://dati.comune.roma.it/download/popolazione-e-societa/popolazione-iscritta-anagrafe-sesso-e-singolo-anno-di-eta-anni-2006
Marriages	Torino	Marriages and spouse's ages, 2011	http://www.comune.torino.it/aperto/dati/demografia/matrimoni-ed-et-degli-sposi-anno-2011.shtml
	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