Semantic Web 0 (0) 1
IOS Press

Open Data Quality from a Developers' Perspective

Antonio Vetrò, Ertugrul Bircan Copur [a], Marco Torchiano [b]
[a] *Technische Universität München*
Garching bei München, Germany
bircan.copur@tum.de, vetro@in.tum.de
[b] *Politecnico di Torino*
Torino, Italy
marco.torchiano@polito.it

Abstract. *Context*: Developers are among the most active consumers of Open Data, building new services and applications upon them. However, often data quality problems limit the potential for this type of Open Data reuse. *Objective:* We aim at understanding if a metric-based evaluation of the quality of Open Data is able to predict the problems experienced by developers building applications that use Open Data. *Method:* We collected from developers the negative and positive aspects of a sample of datasets they used to develop applications, and compared them with the evaluation provided by a set of metrics. *Results:* The main gap between the developers' feedback and the adopted metric-based evaluations was the inability to compare the entities in the datasets to real life references and to detect format problems. We observed a few agreements between developers' perception in Accuracy and Understandability. In addition, from a higher perspective, developers lamented the lack of feedback channels between users and publishers and lack of search mechanisms. *Conclusions:* Although the small sample of datasets and participants used in this study cannot lead to any generalisation, these first results give proper indications on the tuning of the measurement framework to better address developers' issues.

Keywords: open data, data quality, developers behaviour

1. Introduction

The Open Knowledge Foundation (OKFN) defines open data as "data that can be freely used, reused, and redistributed by anyone –subject only, at most, to the requirement to attribute and share-alike". The main idea behind this perspective is that when the data has a wider circulation than the normal unavailability to the public, more interesting reuses of said data could be achieved [1]. Considering the above, legal and technical openness of datasets is not sufficient, by itself, to foster a healthy reuse ecosystem.

Endeavours to increase meaningfulness and reusability of public sector information also mandate representing and managing data so that they can be easily accessed, queried, processed, and linked with other data.

Developers are among the most active users of Open Data, because they build new services and applications upon them. The developers belonging to the organisation that produced the data may rely on implicit knowledge that makes the reuse effective. However the general population of potential reuses have no access to such knowledge. In order to be effectively reusable by any developer, Open Data need a certain level of quality, both in the way they are presented and in relation to their potential usage goals.

In past and ongoing work [12] the authors suggested that a widely adopted data quality model for Open Data, and a set of actionable metrics are useful tools to achieve data quality improvement and, in turn, harnessing the latent potential of Open Data [6] by enabling developers to easily reuse them in their applications. With such a perspective, we have built

an evaluation framework [12] based on the analysis of the methodologies for data quality measurement documented in literature. We defined an initial set of metrics on specific data quality characteristics, which were selected and refined from literature. Eventually we used the resulting evaluation framework to compare the quality of a sample of Open Datasets.

In this paper we move a step forward in our investigation: we report the outcome of an initial validation of the proposed framework. In particular, the aim is to understand whether the conceptual framework and the defined metrics are able to capture the actual quality problems perceived by open data end users, and which problems cannot be identified by a metric-based approach.

With this objective in mind, we reached out to a few developers who worked or are currently working with German open government data to integrate them into their applications, and asked them to evaluate the datasets they have worked on in terms of quality. Then we independently assessed the quality of the relative datasets according to our set of metrics. Eventually we compared the results of the metrics based assessment to the developer reported problems.

In the rest of the paper we briefly discuss previous works in the field of data quality measurement with an eye on the peculiarity of Open Data (Sec. II), we give an overview of the study (Sec. III) and the methodology used (Sec. IV), then present the results (Sec. V) and discuss them (Sec. VI). We report the main limitation of the study in Sec. VII, and finally conclude and explain future work in Sec. VIII.

2. Background and motivation

The attention to Open Data quality has risen over the recent years. One of the best-known works in this field belongs to Tim Berners-Lee, who proposed a deployment scheme entitled "5 stars open data" [3]. This deployment scheme consists of five incremental quality requirements that are represented as stars. While this scheme indeed expresses one of the aspects of data quality, it focuses only on this one aspect, the format used to publish the data; thus cannot by itself be used to assess the total quality of a dataset. In 2007, a more all-around set of principles was produced by a group of Open Data and Internet experts who gathered under the moniker "Open government working group". The original set of principles contains eight rules in total, which state that any Open Data must be: Complete, Primary (as collected at the source), Timely, Accessible, Machine processable, Non-Discriminatory (available without registration), Non-Proprietary (in terms of format) and License-free. The original list has since then been extended with seven more rules, stating that the data must be: Online and free, Permanent (at a stable Internet location indefinitely and in a stable data format for as long as possible), Trusted, Documented, Safe to open, Designed with public input and there must exist a Presumption of Openness [13]. These principles have laid the basis for the development of an assessment process for Open Data quality.

Several data quality models and methodologies have been presented in literature, which were collected by Batini et al. [2], in relation to web portal data quality assessment. In addition to their models, the Software Quality Requirements and Evaluation (SQuaRE) model [7] and Portal Data Quality Model (PDQM) [4] have been developed. A combined model that developed by Moraga et al. titled SQuaRE-aligned Portal Data Quality Model (SPDQM) is later introduced [9], and has been selected as a reference for our empirical evaluations in previous work [12], as it provides a wider set of data quality characteristics then the others. The SPDQM contains 42 characteristics (30 from PDQM, 7 from SQUARE, 5 characteristics were added after a systematic literature review), which are organised in two viewpoints and four categories:

- Inherent viewpoint:
  - **Intrinsic** This denotes that data have quality in their own right
- System dependent viewpoint:
  - **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

Since Open Data typically span heterogeneous domains and they are subject to the most diverse usage from their consumers, 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) in comparison to the other models listed by Batini et al. [2].

This work focuses on the intrinsic quality properties. In a previous work [12], we identified a list of 14 met-

Table 1
Metrics

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 standardised 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.

rics on seven intrinsic quality characteristics to evaluate the quality of Open Data in a few Italian municipalities. The metrics are summarised in Table 1, for the implementation details please see [12]. The metrics we provided are at the lowest possible granularity level, which is cell (according to tabular representation) or dataset level when otherwise not possible. In contrast, similar works in literature assess the quality of Open Data only at portal level (see for example [11] and [8]).

In this work, we address the internal and construct validity of the defined metrics, comparing the output of measurements on Open Datasets with the quality experienced by a small pool of developers that used them.

The motivation for this narrow focus has to be understood in the light of our specific perspective. We, as software engineering researchers, are interested in the quality characteristics – or lack thereof – of the open data sets that enable the development of data-oriented applications.

With such a goal in mind, it is fundamental to provide developers as well as project managers with a tool capable of detecting potential issues that might hinder the effective use of open datasets.

While the metrics defined with a system dependent viewpoint should be able to identify precise problems with the data, precisely because they are system dependent they need to be redefined for every application, consuming both time and effort.

The advantage for a set of intrinsic – system independent – metrics consists in the capability of quickly applying the metrics to every application and obtain an immediate, albeit possibly approximate, assessment of the data quality. This features allow a rough estimation of the potential data reuse and integration problems.

3. Study Design

The goal of this work is to understand limitation and possibilities of a set of metrics on the intrinsic quality of Open Data quality, with a focus on the quality characteristics defined in [12] and from the developers' perspective.

The research questions that guide our study are the following ones:

RQ1. *What are the differences between negative and positive aspects identified by the quality metrics and those experienced by developers?*

RQ2. *What other quality problems, which were experienced by developers, are not detectable with a metric-based approach?*

We aim at providing an answer to the above research questions by means of an exploratory study, which consists of three phases.

In **phase I**, we conducted structured interviews with four developers who used Open Data in their applications. We asked them about the problems they experienced, the relative possible causes, and on the positive aspects of the datasets. The questions are listed in Table 2, with identifier P1-Qx.

We defined six questions to be used as guidelines during the interviews. The interviewees were asked:

- to identify a couple of problems and a positive aspect, and then
- to point out the quality feature that might have caused the problem.

The two questions types were applied for up to three datasest.

In **phase II**, we measured, on the same datasets that have been used by the interviewees, the 14 metrics defined in [12]. The metrics focus on the following six different intrinsic quality characteristics: Traceability, Currentness, Expiration, Completeness, Compliance, Understandability, Accuracy (the definition of the characteristics are taken from [9]).

The outcome of such measurement was then compared the positive and negative aspects reported by the developers. The objective being to understand which problems could have been detected by means of the metrics framework (RQ1).

In **phase III**, we went back to the interviewees to seek for explanations for the differences found between phase I and phase II and gather insights on additional problems identified by developers (RQ2). The questions used in this phase are also listed in Table 2, with identifier P3-Qx.

Overall the design is based on the conceptual model described by the UML [5] class diagram shown in fig. 1.

Every Developer participating in the Phase I interviews reports different aspects concerning the datasets he/she worked with; such aspects can be either negative (P1-Q1/Q3) or positive (P1-Q5). Such aspects concern different quality characteristics (P1-Q2/Q4/Q6).

The same dataset are evaluated, with respect to the same quality characteristics by collecting measures of the metrics defined in the adopted framework. In the framework there may be more than one metric per characteristic.

4. Analysis Methodology

We analyse the results of phase I and II by comparing the positive and negative aspects that emerged in the two phases.

To analyse the results of Phase I we computed the proportion of participants that mentioned a problem (or a positive aspect) related to each quality characteristic:

(a) Accuracy
(b) Completeness
(c) Compliance
(d) Currentness (corresponding to Actuality in Tab. 2)
(e) Traceability

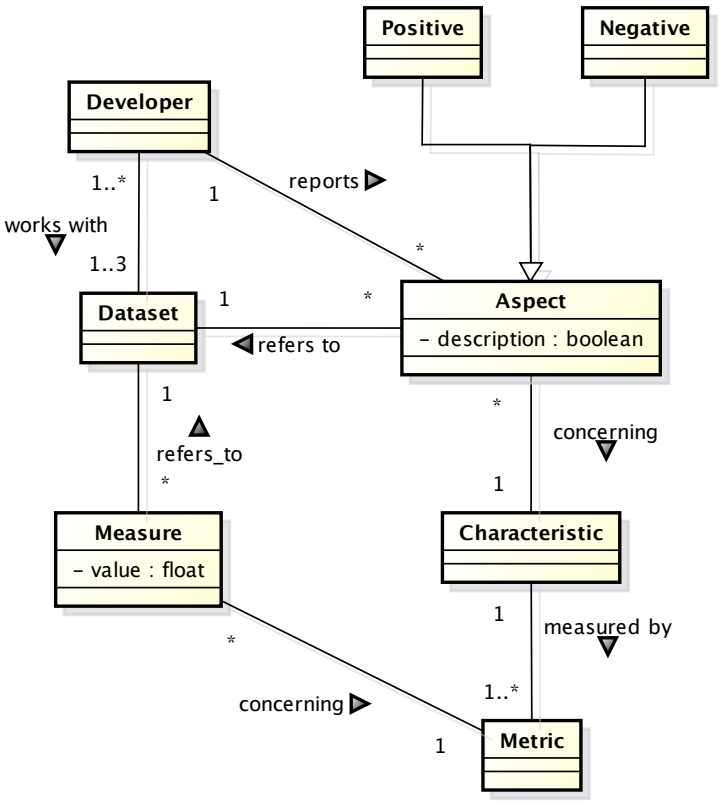

Fig. 1. Conceptual Model

(f) Understandability (corresponding to Metadata in Tab. 2)

We consider a quality characteristic as potential problem (-) for a dataset when the number of developers that reported a problem referring to the dataset concerning the given characteristic is greater than those reporting a positive aspect.

Similarly we identify the developer-reported potential positive links (+) between quality characteristics and datasets. For the dataset the were never related to a quality characteristic, we assume a neutral relationship (0).

In addition we take into account when problems or positive aspects reported in open questions (P1-Q1/3/5) could be linked to other quality characteristics, and in that case we assign, respectively, a (-) or (+) .

For phase II, we measured the selected metrics on the very same datasets reported by the participants interviewed in phase I. Metrics are normalised to the interval (0;1), where a value of 1 represents high quality and a value 0 represent the lowest possible quality level (metric implementation details can be found in [12, p 17 and Appendix B])

In order to map the metric value to a (-) or a (+) , we have to take into account the fact that we are using more than one metric to evaluate each characteristic. For example Understandability is measured by both "Percentage of Columns with Metadata" and "Percentage of columns in an understandable format". In one of the datasets the first metric was zero, as there was no metadata provided, but the second was 0.85. In this case the total result can neither be considered a good one nor a bad one. For this reason we applied the following methodology: when two metrics for the same characteristic yield such different results, we counted the evaluation as positive for average > 0.6, and as negative for average < 0.4. The $0.4 - 0.6$ interval is considered neutral. This holds for all datasets and characteristics.

In phase III, we coded the transcripts of the interviews following standard procedures specified in the literature [10]. Since the interview was created specifically for the purpose of finding the most problematic aspects of the datasets according to user responses in

Table 2
Questions used for the interviews conducted in phases I and III.

Phase	ID	Question
I	P1-Q1	Please describe the **main problem** you have encountered in the dataset, regarding the quality of the Open Data, during the development of the application. (For up to three datasets)
	P1-Q2	Please indicate which characteristic(s) of the dataset might be, in your opinion, the cause for that problem, among the following ones: *(a) Accuracy (b) Completeness (c) Compliance (d) Actuality (e) Metadata*
	P1-Q3	Please describe the **next main problem** you have encountered in the dataset, regarding Open Data quality, during the development of the application. (For up to three datasets)
	P1-Q4	Please indicate which characteristic(s) of the dataset might be, in your opinion, the cause for that problem, among the following ones: *(a) Accuracy (b) Completeness (c) Compliance (d), Actuality (e) Metadata*
	P1-Q5	Please describe the **main positive aspect** you have encountered in the dataset, regarding OpenData quality, during the development of the application. (For up to three datasets)
	P1-Q6	Please indicate which characteristic(s) of the dataset might be, in your opinion, the cause of this positive aspect, among the following ones: *(a) Accuracy (b) Completeness (c) Compliance (d) Actuality (e) Metadata*
III	P3-Q1	Has the lack of metadata been an issue for you? When the metadata were not present, were the column names sufficiently explanatory?
	P3-Q2	Have you ever encountered a case where upon finding an empty field in the dataset you could not decide whether that field was intentionally left blank, or the data was simply missing? If you did, how do you think a distinction can be made between those two cases?
	P3-Q3	Can you easily find the data you are looking for among different data portals in Germany? If not, how do you think this process could be improved?
	P3-Q4	Were you able to find the data you needed in a suitable format? Were there any problems in the data format, or was it easily usable?
	P3-Q5	In your opinion, what is the one thing that definitely needs improvement regarding open government data?

the phase I, responses were just tagged with the categories previously defined. The ideas and feelings of the users about the datasets and the open data publishing processes were taken objectively as is, and are reflected as detailed feedback to data publishers in terms of most common problems from the open data community.

5. Results

In the first phase we interviewed four developers, which reported a total of six distinct datasets that they have worked with (however one of these datasets was not in an usable format for us, as it was a live API call with constantly changing values, while the metrics framework we adopted can be applied to tabular data only). Most of developers developed apps for smartphones, and many of them made use of geographical data. The technologies they used are heterogeneous: HTML, JS, CSS, PHP, Leaflet, Python, Objective-C, MapKit, CCHMapClusterController, MongoDB. The level of experience in programming and using Open Data was at least five years for each person (except one which had less then 5 years experience with Open Data). The datasets analysed are the following ones:

- Dataset 1: List of used glass containers in the Charlottenburg - Wilmersdorf area[1]
- Dataset 2: List of popular first names in the Berlin area, in 2013 [2]
- Dataset 3: List of Christmas markets in Berlin, in 2014 [3]
- Dataset 4: List of bus stops in Berlin, in 2012 [4]
- Dataset 5: List of memorial stones in Berlin [5]

We report the comparison between the results from interviews with developers (D) and the metrics (M) on Table 3. On the columns, in correspondence to quality characteristics, we indicate with the sign "+" a characteristic observed as positive, similarly with the sign "–" a characteristic perceived as negative, while the sign "0" marks characteristics where the participant or the metrics measurements did not indicate any clear

[1] http://goo.gl/SMW9qM (last retrieved: 22 June 2015)
[2] http://goo.gl/oxTAQb (last retrieved: 22 June 2015)
[3] http://goo.gl/IEThNT (last retrieved: 22 June 2015)
[4] http://goo.gl/zYBFCQ (last retrieved: 22 June 2015)
[5] http://goo.gl/jtdWaW (last retrieved: 22 June 2015)

Table 3
Results of Phase I and Phase II

Dataset	Completeness		Traceability		Compliance		Accuracy		Understandability		Currentness	
Phase:	D	M	D	M	D	M	D	M	D	M	D	M
1- Used Glass Containers	0	+	0	–	–	–	0	+	0	0	0	n/a
2- Popular First Names	–	+	0	+	0	+	0	+	0	0	0	n/a
3- Christmas Markets	–	+	0	+	0	+	–	–	0	0	–	n/a
4- Bus Stops	+	–	0	+	0	+	+	–	–	–	0	n/a
5- Memorial Stones	–	–	0	+	0	+	–	–	–	0	+	n/a

Table 4
Results of phase III

ID	Summary
P3-Q1	Lack of metadata Often undecipherable column name or values
P3-Q2	*Meaning of empty cells unclear*
P3-Q3	Too many data portals in Germany No standardisation between portals
P3-Q4	Problems with data formatting and localisation Data not always available in all open formats
P3-Q5	*Lack of a proper communication channel between open data users and publishers*

trend. When a metric was not applicable to a dataset, the result is "n/a". The rows correspond to the different datasets analysed. We color in red the cells corresponding to a disagreement between metrics and developers (positive versus negative evaluations), in green when there is an agreement (both evaluations positive or both negative). In all other cases we keep a conservative approach and we do not infer any conclusion from the data.

Finally, the results of phase III are summarised in Table 4, where we report for each question the key points emerged.

6. Discussion

The first phase participants interviews highlighted several problems with the quality characteristics of the datasets, *Completeness* being the most problematic. On the contrary, our metrics show that completeness is one of the positive aspects of the datasets. The meaning of this difference is that the metrics refer to Completeness as ratio of not-empty cells, while the interviewees had also in mind how complete the datasets were in comparison to the real world entities references in the datasets. Further insights concerning Completeness and the problem of empty fields came from phase III (P3-Q2): both interviewees reported that this problem creates the perception in the user that the dataset lack of completeness. As potential solution, they suggested that null values could be used in appropriate fields to disambiguate possible interpretations, in conjunction with use of metadata.

We observe in Table 3 a second discrepancy between participants' answers and metrics for *Accuracy* in dataset 4: however, there is agreement in datasets 3 and 5.

We do not observe other disagreements between developers' feedback and metrics results, but one negative agreement for *Compliance* in dataset 1. Regarding this aspect, also, we gathered more understanding from phase III. According to the interviewees (P3-Q3) Open Data portals in Germany are not standardized: the formats of data they present to the user interfaces are very different from one another. Problems on data format were reported also in (P3-Q4): for instance, both our interviewees reported that in sources with GeoJSON formatted values, the values they could gather were not parseable without any modification. The most common problems they encountered included the false usage of commas and dots in numbers, and character encoding problems. Also we discovered that the characteristics and metrics chosen for the framework are not able to detect redundant values in the dataset, mostly important duplicates.

Concerning *Understandability* and especially the metadata (Understandability, Question P3-Q1), the interviewees said they sometimes encountered abbreviations in columns that they did not know how to interpret, and metadata was not able to guide them.

To summarise, we conclude with the following answer to RQ1: The main difference between the metrics and the perception of the developers regards Completeness. Despite agreement, the metrics were not able to capture problems regarding compliance of data formats. In addition, a common problem

Regarding RQ2, interviewees reported the lack of a search mechanisms: often the best way to find a certain dataset from a given city or state was usually to ask personal contacts (P3-Q3). Question P3-Q5 in our interview showed that another problem with the current Open Data publication scheme is the lack of proper feedback channels. Right now the way the feedback mechanism works is the users sending an email or filling a contact form in the portal, and not knowing whether they will receive a response. Our interviewees think that the lack of such proper channels reduces liability on the publishers' part, and leads to lesser maintenance of datasets.

In summary, our answer to RQ2 is: *The main quality problems not detectable by a metric-based approach are the lack of an efficient way to search for open data sets and the lack of proper feedback channels between Open Data users and publishers*.

7. Threats to Validity

As it was not possible to ask users to comment on every aspect of a dataset, our investigation with its low number of participants does not aim neither at completeness nor at generalisation.

In addition we might have missed or misinterpreted the link between problems and positive aspect to quality characteristics.

8. Conclusions and Future Work

This exploratory study gives us initial indications on limitations and opportunities of using a metric-based approach to evaluate the quality of Open Data from the developers' perspective. Our next necessary goal is to improve our sample set and make these evaluations on a wider scale. In the long run, we aim to derive a set of practical guidelines and an automatic tool for Open Data publishers, to complement the existing recommendations on Open Data disclosure.

Acknowledgments The authors are thankful to all participants to the study.

References

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

[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), July 2009.

[3] T. Berners-Lee. Linked data-design issues. Technical report, W3C, 2006.

[4] A. Caro, C. Calero, I. Caballero, and P. M. A proposal for a set of attributes relevant for web portal data quality. *Software Quality Journal*, 16(4):513–542, 2008.

[5] M. Fowler. *UML Distilled: A Brief Guide to the Standard Object Modeling Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 2003.

[6] J. Hofmokl. The internet commons: toward an eclectic theoretical framework. *International Journal of the Commons*, 4(1):226–250, 2010.

[7] ISO/IEC. 25012 international standard: Systems and software engineering - software product quality requirements and evaluation (square)-data quality model. Technical report, ISO/IEC, 2008.

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

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

[10] C. Seaman. Qualitative methods in empirical studies of software engineering. *Software Engineering, IEEE Transactions on*, 25(4):557–572, July/August 1999.

[11] B. Ubaldi. Open government data: Towards empirical analysis of open government data initiatives. Technical report, OECD Publishing, 2013.

[12] A. Vetrò, M. Torchiano, C. Minotas Orozco, G. Procaccianti, R. Iemma, and F. Morando. An exploratory empirical assessment of italian open government data quality with an eye to enabling linked open data. Technical report, Politecnico di Torino, 2014.

[13] VV. AA. The annotated 8 principles of open government data. the 8 principles of open government data, 7 additional principles. Technical report, 2014.