HAL open science

Towards Open Data Quality Improvements Based on Root Cause Analysis of Quality Issues

Csaba Csáki

To cite this version:

Csaba Csáki. Towards Open Data Quality Improvements Based on Root Cause Analysis of Quality Issues. 17th International Conference on Electronic Government (EGOV), Sep 2018, Krems, Austria. pp.208-220, 10.1007/978-3-319-98690-6_18 . hal-01961523

HAL Id: hal-01961523

https://inria.hal.science/hal-01961523v1

Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Towards open data quality improvements based on root cause analysis of quality issues

Csaba Csáki[1][0000-0002-8245-1002]

[1] Corvinus University of Budapest, Budapest, Hungary
Csaki.Csaba@uni-corvinus.hu

Abstract. Commercial reuse of open government data in value added services has gained a lot of interest both as practice and as a research topic over the last few years. However, utilizing open data without proper understanding of potential quality issues carries the risk of undermining the value of the service that relies on public sector information. Instead of establishing a data quality assessment framework this research considers a review of typical open data quality issues and intends to connect them to the causes leading to these various data problems. Open data specific problems are concluded from a case study and then theoretical and empirical arguments are used to connect them to root causes emerging from the peculiarities of the public sector data management process. This way both practitioners could be more conscious about appropriate cleansing methods and participants shaping the data management process could aim at eliminating root causes of data quality issues.

Keywords: open data, open data reuse, data quality framework, open data quality, public procurement, data cleansing, root cause analysis.

1 Introduction

While the idea of open government data (OGD) has been around for some time, every few years there is a rejuvenated scientific interest in the topic. The motivation had been changing from promoting accountability and transparency to supporting e-Government to the push for open government. The latest trend is based on economic interest, namely the idea of innovative, commercial reuse of public sector information (PSI) [14]. The changing focus came with changing research goals and shifts in research objectives.

One of the main conditions of successful reuse is quality data [6]. It is no surprise, therefore, that over the years a lot of scientific and on-the-field efforts have been reported to address the issue. One well-quoted are deals with the question of provenance [21] with the leading idea covering the maturity level of linked open data published [2]. However, the shortcoming of the five star model is that it focuses almost exclusively on the area of linked open data, thus only addresses the related subset of quality features (such as traceability, linking standards, and machine readability). There are ample results that offer frameworks dealing with a wider range of quality dimensions – both in general [32, 17] or focusing on OGD in particular [10] –, but

these approaches mostly cater for the assessment of datasets [1, 30, 35]. While the ability to judge the quality of a given data set is useful, typically there are no guidelines how to address these issues – neither how to eliminate them on the producer side, nor what to do about them before (secondary) utilization. While considering the outcome could improve certain aspects of structuring and storing data, existing frameworks are unable to consider the bigger picture of policies and regulations or the organizational context [36]. So, although there are ample proposals about organizational preparation or guidelines considering technology, even these best practices do not seem to be enough regarding quality of the content [37]. The assumption here is that guidelines related to various aspects of quality are no help mainly because they do not focus (enough) on the root causes of ODG quality issues. Thus while there is considerable research efforts addressing certain specific areas related to producing good quality open data, relatively less attention is paid to why data quality issues are still present despite the best efforts and available experience.

The research reported here was focusing mainly on the latter problem and its objective was to understand the root causes of quality issues affecting the publication of PSI intended for reuse. This paper is organized as follows: the next section reviews frameworks of open data quality and considers characteristics of the public sector as context. The methodology section raises research questions and presents the methodology followed. Then the attention is turned to a review of public sector data quality issues through a case study. This is followed by a discussion of potential root causes of issues identified along with their generalization. The paper closes with conclusions and the usual recommendations of future research ideas.

2 Open government data and related quality considerations

2.1 The push for open government data

The term open data as a popular concept was first used in 2006 when the Open Knowledge Foundation (OKF) has announced its Open Knowledge Definition [18, 5]. This was a general call for opening up scientific and other data for further use or reuse. There are other forces behind publishing public sector data, however, some of which are rooted in early civil society movements, albeit the dawn of the Internet and related technologies have also played a major role in an increased interest (by providing new possibilities). Meeting integrity and accountability goals in democratic societies is anchored by transparency that in turn assumes a requisite level of openness whereby non-government actors (the public) have mechanisms to know what governmental actors are doing [4]. Thus, data about governmental behavior may be used to hold actors of the public sphere to account [20]. Over time the goal of promoting accountability and transparency has been overtaken by the motivation to use technology in support of e-Government initiatives [12], and later by the push for open government [34] also fueled mostly by advances in technology. In the former technology was an enabler of open data, while in the context of open government open data became the main piece of the puzzle as an enabler of advanced participative democracy [28]. However, the latest push for publishing data generated or controlled by public

actors considers commercial utilization of such data. Indeed, reusing public sector information in innovative value added services is turning into a serious market [19].

Since more and more data and information generated in various public policy domains are being captured, digitized, and stored, it would be difficult (at least in democratic regimes) to argue for completely shielding such digital records from public scrutiny on the one hand or potential utilization on the other. While users of public or open government data are assumed to be able to utilize them with ease, the fact is that open data is not free of quality issues [3]. The higher demand for more open data does not necessarily come with increased quality. Assessing the quality of data in general or open data in particular is not a straightforward exercise.

2.2 Information quality dimensions

There are different frameworks allowing for discussion over data or information quality (DQ/IQ) and within which DQ/IQ may be assessed (for simplification we do not enter into a theoretical discussion about the difference of data and information, we simply regard information as data in use, as that fits the problem at hand). The technical view associates quality with the accuracy of the data in products such as databases looking at timeliness of update, system reliability, system accessibility, system usability and system security [16]. Another, the machine readability approach [35] is concerned with linking, finding, relating and reading data typically using automated processes, and characteristics usually considered include number of formats, traceability, automated tracking, use of standards, or provenance. Perhaps the most commonly used simple definition of user side information quality interprets the term as "fit-for-use" [31]. However, IQ defined this way remains a relative construct whereby data considered appropriate for a given use may not display acceptable attributes in another setting [25]. Furthermore, fit-for-use does not immediately allow for ready measurability and it requires additional detail in order to be operationalized [11].

When it comes to actually assessing DQ/IQ, it is typically related to "a set of 'dimensions' that are usually defined as quality properties or characteristics" ([22] p. 2). However, this approach has led to a proliferation of features and dimensions – as various models proposed distinguished sixteen [15], twenty-eight [7], or even thirty-two [29] different dimensions – although the most important dimensions appear to overlap. It is possible to organize these features along a natural timeline of the steps normally taken when exploring new data: Awareness and availability, Accessibility, Readability, Technical qualities of the data, Content and structure, Traceability, Usability, Fit-for-purpose – each in turn including several sub-features (depending on the framework). However, the quality of the data as stored, accessed and manipulated can substantially differ from the quality of the information that the data contains or that the data can offer in terms of information gleaned from it.

2.3 Some special characteristics of the public sector

As an important step towards understanding why quality issues may happen in public context, it is worthwhile to look at a few fundamental characteristics of the public

sphere. One essential difference between the private and the public sector is that public policies are created based on public interest while private corporate goals serve private interest. Public value cannot be defined by commercial categories only, and governments thus have responsibilities related to fulfilling non-commercial goals which in turn increases costs [9]. In addition, governmental choices have lasting, long-reaching effects. Indeed, the policy making function creates the – formal and legal – environment within which society and economy operates. [23] points out that the notion of public interest does require some form of sympathy with the needs of others which, therefore, may not be reconciled with the market notion of maximizing economic opportunities and personal wealth.

Regarding data management, governmental functions are guided by laws and regulations influencing related processes, tasks, roles, and responsibilities. Furthermore, there are specific regulations controlling the release of data (typically in the context of the so called 'right to information' law). Thus, utilizing data from public websites presumes some level of legitimacy on the part of the immediate publisher. This setting has an impact on the way of producing and collecting data as well as the way open data is generated from the data stored. Information collection is usually done through forms defined by the corresponding law such as a relevant act. All in all, these result in a data lifecycle that is different from its counterpart in the private sector: the data does not connect, rather open data as published sharply separates the supplier from the consumer. In addition, changing any part of the data producing side would require changes in the corresponding regulatory component - which might take time (due to the formal processes involved).

2.4 Open government data quality frameworks

Despite the difference in context, open data specific ‘quality frameworks’ often offer similar categorizations. One ODQ stream considers technical standards and abilities as well as processes and outcomes of producing and managing datasets, but also considers the timeliness of data (i.e. whether it is out-of-date). Another stream is centered on the availability and accessibility of various types of data or data in certain categories, while also measures whether intended audiences are aware of the availability of relevant datasets and if data is easy to find [5]. Yet another set of frameworks is concerned about specific sectors and take into account the content of the datasets. Finally, it is customary to ask about the value of open data, which, in general terms considers the needs of end users [11]. However, irrespective of the approach, there is a tendency here too to favor characteristics that are measurable. The current disposition of open data quality characteristics is aptly demonstrated by [35] which, in pursuit of the measurability of ODQ, define and operationalize 68 metrics along 6 dimensions. Most of these characteristics do not differ from ‘regular’ data quality dimensions, although there are some additional concerns related to access to and freshness of data – as well as to potential fees charged. However, irrespective of the assessment methods, there are still typical problems with public data made available. Therefore, the objective of this research was to understand and map the root causes why (and where) certain

types of public open data quality issues and errors happen – with the ultimate intent to make recommendations what to do about them.

3 Methodology: Theoretical arguments with a case study

Under the above objective the following research questions are proposed: 1) What is the status of the quality of Open Government Data from the point of view of 'content'? 2) What are the main reasons (root causes) behind OD content issues in public context? 3) What to do about improving content quality of OGD? This paper focuses on the second question, using the first as support – but would not have the room to address the last question here.

To demonstrate typical content errors in public sector data sets along with their causes a case study methodology was designed [33]. To explore the reasons behind the quality issues of open government data the research plan contained the following steps: 1) review typical data quality issues (that impact reuse); 2) illustrate them through examples from the case; 3) use the case study to identify potential causes of errors; 4) propose a generalization of those root causes.

This was an intrinsic [24], single case [8] research study, where the exploration of the case (i.e. data collection) involved a) investigating a complex open data set; b) reading documents describing the data set (including its structure and known issues); and c) email communication with a representative of the issuer of the data for further clarification. To establish root causes of issues presented in the case theoretical arguments from relevant literature were applied. Finally, in the last step the generalization was based on the understanding of the immediate context and process of producing open government data (concluded from the literature review).

The data set used (as the case) was the public procurement (PP) open data of the European Union (EU), selected based on its special characteristics regarding size, complexity, regulatory context and multiple stakeholders. Counterarguments may be that the EU PP legal context is complex and further burdened by a multinational setting. However, while understanding the depth of the case might be a challenge for some readers outside PP, the descriptive power of the case well offsets the efforts required. As part of its broader e-Government initiative the European Commission (EC) has been an advocate of the open data movement for some time. "The European Data Portal" (https://www.europeandataportal.eu/) offers public sector information originated in the member states and portal data may also be repurposed. Through the Directive 2003/98/EC the EC has set up the legal framework to allow the reuse of public sector information. One key component of the EU Open Data initiative (https://ec.europa.eu/digital-single-market/en/open-dat) is the Tenders Electronic Daily dataset comprised of public procurement data of the twenty-eight member states. While the data is accessible as part of a daily journal (the online version of "*Supplement 32 to the Official Journal of the European Union*"), there is an annual release of summarized historical data in CSV format (dating back to 2006 at https://data.europa.eu/euodp/en/data/dataset/ted-csv). Data covers purchases of public procurement that fall above given threshold amounts stipulated in EU regulations for

procurement. Other than EU members, affiliated countries also publish tender and award notices in the TED Journal to gain access to the EU market. Data in the Journal are collected from standardized public procurement forms as required by the corresponding EU Directives (2014/17 and /18) and their Annexes. The data originally recorded store information captured from the contract notices reported in standard forms #2, #4, #5, or #17. These forms announce information concerning a future purchase (i.e. call for tender). Another set of data covers contract award notice information on the outcomes of the procurement obtained from standard forms #3, #6 or #18. Data is entered through online versions of these forms, one notice at a time. The open datasets published annually come with a codebook [26] describing the fields in the files made available. In addition, for advanced users of the CSV datasets a user guide is available [27] providing information about known issues and difficulties.

The TED open data is very complex because the CSV data files have three levels of procurement information embedded: a) contract notices (CN); b) contract award notices (CAN); and c) contract awards (CA) (the last two published in one file, of course). While the process of public procurement is inherently complicated, for now it should suffice to state that one and occasionally two CNs lead to one CAN (this is because a CN may have a preliminary notice with a separate CN ID), but one CAN may lead to one or more CAs associated with it (as a single call may have several parts or lots with each leading to a separate contract being awarded under the same CAN ID but individual award CA IDs). Issuers of notices are called "contracting authority". Each annual dataset is published in CSV format using UTF-8 coding. All data files were (first) downloaded January 17, 2017. There were two types of data – notices and awards – from 2009 to 2015 (the first three years had to be omitted), and the size of the sixteen files was over 2 GB (each ranging from 130 to 280 MB). MS Excel and MS Access (both from Office 2010 on Win7 OS) have been utilized to open and investigate the structure and content of the files. In addition, SPSS (v22.2) and Oracle Database (11g r. 11.2.0.4) were also used to investigate data quality.

4 Typical quality issues in the case – and their root causes

While access related features (Availability, Accessibility and Readability as well as Traceability) are important as a starting point for OD utilization, they are less relevant in our context of actual reuse. On the other hand, content related characteristics (such as Technical qualities, Content and structure, as well as Fit-for-purpose) are main concerns that immediately influence usability. According to literature, data content errors are typically organized into four categories (based on [25]): Missing data (missing field or missing value); Duplication (physical duplication or logical duplication); Error with meaning or interpretation (syntactical error, out of bounds, format error, data does not make sense in context); and Inconsistency (inconsistency between data fields, data tables, databases or outside sources).

Case problem: Successful opening of the file(s) is followed by the investigation of the Technical qualities of the data. Due to the nature of CSV, the original dataset as published does not carry datatypes. The typical result is formatting errors. Even in

Excel – the tool users would use to open CSV –, fields containing data that look like (calendar) date would indeed be interpreted as calendar date, resulting in automatic corrections, which are often faulty (e.g. 2004 may become 2004 January 1).

Root cause analysis: These issues are related to the process of generating open data and the publishing format being used. For example, a formatting error may be the result of inadequate consideration and lack of flexibility in data formats (especially in international, multi-language context). Also, lack of data type information in simple standard formats may lead to misinterpretation by more sophisticated tools.

Case problem: Since EU members may use any of the official languages for their PP announcements, basic UTF-8 reading with a default language (such as English) resulted in scrambled characters for languages like Greek, Hungarian, Swedish, etc. Interestingly, each tool used had its own way of dealing with this problem: SPSS, and Oracle could only read setting of one language or another, MS Access had an "all" setting for UTF-8 font mixing, allowing for text from every EU language to be displayed properly, while MS Excel required the "import" function for proper UTF-8.

Root cause analysis: It appears that CSV does not carry language information, but UTF-8 requires a so called BOM character for font mixing.

Case problem: The case data files had a lot of text fields, some of them are quite long – and for most tools the length of textual data is an issue: some truncate lines while others simply drop whole records with fields of inappropriate sizes. So, as the result of the above, the actual data as opened may have missing fields, missing content, or inappropriate content or even inconsistencies within the dataset.

Root cause analysis: The loss of information is due to technical issues such as the lack of data type information or the use of long text in one field.

Case problem: The most important step (before any use), is the checking of content and its validity. The outcome of a procedure (CN) may be a successful award (CAN with one or more CA), modification, cancellation, or cancellation with a new call. Unfortunately, cancellation and modification information are not always recorded properly (cancellation or modification flag is missing from the form) leading to either missing information or duplicate records. As a result, the CSV output generated had missing flags and duplicate CN IDs.

Root cause analysis: The cause of such problems is rooted in the mode of entering data into the forms, especially online: a) the forms themselves could be faulty (such as having missing fields); b) there could be human error (using the wrong form or lack of knowledge about how to fill out the form) on part of the contracting authority personnel; and c) these may be combined with inappropriate sanity check or lack thereof. In addition, d) the algorithm generating the OD output file may be misled by the inappropriate information.

Case problem: Another form of information loss happened when there were multiple values in one field and most tools could not separate them. This happened in two ways: a) when there were two winners to be announced, instead of two separate award (CA) IDs the name of both winners were entered into the corresponding field; b) categorization of the product to be purchased is based on so called CPV codes, but complex purchases may require one main and several secondary CPVs.

Root cause analysis: Situation "a)" is clearly a human error; while "b)" relates to the way forms are defined (instead of allowing for recursion, repeated values are entered into the same field using some separators). The latter issue causes a problem either way when output records are generated.

Case problem: There were duplicate lines where certain CN IDs were erroneously coupled with CAN IDs from other calls. While CANs may appear several times in case of multiple awards for one call, calls (a given CN ID) should not be repeated.

Root cause analysis: This appears to be a CSV generation issue, as checking such duplicates on the TED search page returns only one item for a given CN ID. Furthermore, in 2014 there was a change in forms – and generating the CSV data from data captured using the old forms were executed according to new forms leading to irregular duplications (which could have easily been filtered out).

Case problem: There were inconsistent values: each type of call should use the corresponding form, however, there are a reasonable number of records where the form number in the record does not match the type of call.

Root cause analysis: This is a data validation issue during the submission of the form (likely coupled with human error).

5 Generalization of the causes behind OGD quality issues

It follows from the nature of public sector activities, namely that they are governed by policy (with underlying strategies) and corresponding laws and regulations, that legal foundation for publishing OGD could already have an influence on the data that may, must, or should not be released and how they were supposed to be published [13]. The legal frame controls what may be published, in what format or by whom. This carries a certain risk of errors when it comes to content and format of data being made openly accessible.

Implementation of the regulations poses a challenge as well – organizationally, process-wise and regarding technological support. In the case presented, the EC directives stipulate that the collecting and entering of data is organized around filling out specific forms. These are not always on-line, thus entering data online often means copying from hand-filled forms. This is a major source of typos and errors. Even with online forms there is a possibility of inappropriate completion of data fields – some of which are deliberate [27]. Individual behavior and lack of control mechanisms built-in when uploading data using the forms will eventually lead to error in generating the output format from data stored. Fields may be missing from the form, data is not even entered into the form (field left empty). Even if data was entered, often the data is a dummy value just to fill in the field (to avoid being caught by validation if the field is empty). Allowing multiple values in the same field is a serious form issue, resulting in serious challenges during statistical analysis.

Understanding the meaning of various fields requires in depth knowledge not only of public procurement in general but specific details of EU procedures, including the intent and use of various forms. For example, fields in the csv files did not fully reflect either the fields in the TED DB (presented as documents through an online inter-

face) or the original forms contracting authorities required to use when submitting data related to calls and results. This is not unusual for public sector data collection typically based on forms. Based on the analysis in the previous section and on the understanding of the role of forms in the public data management process, Table 1 summarizes the generalization of root causes identified.

Table 1. Overview of issues and causes – through examples

Type	Case examples	Root cause	Reasons generalized
Format issues	Date is not interpreted properly;	Representation and data type issue;	Either in the DB or during generation of the open version inconsistent formatting is used – and most often data type information is lost;
	Scrambled characters appeared for certain countries;	UTF-8 does require BOM for font mixing;	Machine readability of even standard forms have language dependencies;
Missing data	No indication of cancellation;	Form error;	Public sphere data collection forms are part of the regulation but often are out of sync with the process;
		Data entering error;	Due to the complexity of legally controlled processes, mistakes are easy to make;
		Error in checking the validity of filling out the form;	Checking relationships between data being entered and data in the DB is not straightforward in this context;
Logical duplication	CNs are mixed with CAN belonging to a different procedure;	Output generation error: During the generation of the open version (CSV), records were connected inappropriately;	Data recorded in form (using online of pdf) are then stored in various databases and the open version is generated using a dedicated process (and algorithm) which may introduce errors;
Physical duplication	Two winners announced in the same field instead of using two separate award (CA) IDs;	The form allows for long text fields and it is difficult to detect whether there are one or more winners;	Lining up the process and the forms is difficult – which makes any automatic detection of form errors complicated;
Content error	Long text of purchase data is truncated by certain tools;	No limit on size of text fields;	Forms collecting data allow for lengthy textual information;
	CPV codes may have	Multiple values are	IT is a typical form defini-

	several values in the field;	allowed in one (text) field;	tion error where database representation (and analytical) requirements are not considered;
	Purchase value is not realistic (e.g. € 1234567);	Contracting authorities have no intention to publish certain data;	Deliberate misrepresentation;
Inconsistency	There were calls without cancellation or eventual awards for years;	Inconsistency in using or filling out the forms – as contracting authorities did not submit a cancellation notice;	Forms are complex and they are difficult to change. Furthermore, due to the large volume of the data there is no bandwidth (process or technology) to detect inconsistencies;
	Type of call value does not match the actual form used (or should have used a different form);	Each type of call notice has its own form (#) – but often the number entered into the form is wrong due to human error;	Although there are human errors, often certain basic errors may or may not be detected by the sanity algorithms;

Overall, it can be concluded from the table (and the examples presented in the analysis section), that in the context of the public sector quality of open data is far from being a simple technical issue (or a DB problem). Issues with the official forms mandated, the complexity of the process, or even deliberate misrepresentation of data may hinder the usability of the data eventually published – on top of regular technical challenges of formats, representation, and readability. In addition, the process of generating the data set version intended for open publication may bring in further errors.

6 Conclusions and practical results

This research paper has argued that while current OD quality frameworks are strong tools when it comes to assessing OD quality or measuring maturity of data released, they are inadequate when it comes to helping public organizations how to release their data in better shape and how to improve quality as experienced by the end user during reuse. It was proposed, that an investigation of the root causes leading to lower quality PSI/OGD is needed and the idea and its possibilities were demonstrated through a case study.

It was demonstrated through the causes identified that ensuring quality of open government data is not simply a technical exercise and often even good organizational practices might not be enough. Although proper data governance principles augmented with well-organized data management and release processes could certainly im-

prove, quality starts at the forms and rules set out in regulations. Therefore, for deeper quality improvements changes need to reach as far as the level of policy frameworks.

During the execution of the case study data quality issues identified in the TED csv datasets had been communicated to the issuer of the data. As a result, first the aforementioned "advanced notes" [27] had been released, and later improvements have been made to the production of the TED OD – with latest datasets released during the completion of this paper (i.e. changes could not be included here). In addition, the codebook [26] has been updated as well. An obvious next step is to investigate the content and changes of the new datasets. The publishing team could also be contacted again in order to collect information about the actions taken to improve quality: this could help further validating the root cause analysis presented here, potentially leading to advanced guidelines for issuers of PSI/OGD.

References

1. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys 41(3), 16–52 (2009).
2. Berners-Lee, T.: Linked Data. IJSWIS 4(2) (2006).
3. Bertot, J.C., Gorham, U., Jaeger, P.T., Sarin, L.C., Choi, H.: Big data, open government and e-government. Information Polity 19(1-2), 5-16 (2014).
4. Bovens, M., Goodin, R.E., Schillemans, T. (Eds.): The Oxford Handbook of Public Accountability. Oxford University Press, Oxford (2014).
5. Davies, T.: Open Data Barometer. Global Report 2013. http://www.cocoaconnect.org/publication/open-data-barometer-2013-global-report, last accessed: 2017/0/21.
6. Dawes, S.S., Helbig, N.: Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency. In: 9th IFIP WG 8.5 International Conference on Electronic Government (EGOV). Springer, pp. 50-60 (2010).
7. Dedeke, A.: A Conceptual Framework for Developing Quality Measures for Information Systems. In: Proceedings of the 5th Int. Conf. on Information Quality, pp. 126-128 (2000).
8. Eisenhardt, K.M., Graebner, M.E.: Theory building from cases: Opportunities and challenges. Academy of Management Journal 50(1), 25-32 (2007).
9. Erridge, A., Hennigan, S.: Public Procurement and Social Policy in Northern Ireland: The Unemployment Pilot Project. In: Thai, K. V. and Piga, G. (Eds.), Advancing Public Procurement: Practices. PrAcademics Press, Boca Raton, Florida, pp. 280-303 (2007).
10. Erickson, J.S., Viswanathan, A., Shinavier, J., Shi, Y., Hendler, J.A.: Open Government Data: A Data Analytics Approach. IEEE Intelligent Systems 28(5), 19-23 (2013).
11. Frank, M., Walker, J.: User centred methods for measuring the quality of open data. The Journal of Community Informatics 12(2), 47-68 (2016).
12. Grönlund, Å., Horan, T.A.: Introducing e-gov: history, definitions, and issues. Communications of the AIS 15(1), Article 39 (2005).
13. Huijboom, N., Van den Broek, T.: Open data: an international comparison of strategies. European Journal of ePractice 12(1), 4-16 (2011).
14. Jetzek, T., Avital, M., Bjorn-Andersen. N.: Data-driven innovation through open government data. J. of theoretical and applied electronic commerce res. 9(2), 100-120 (2014).
15. Kahn, B.K., Strong, D.M., Wang, R.Y.: A Model for Delivering Quality Information as Product and Services. In: Proceedings of the 1997 Conference on Information Quality, Cambridge, MA, pp. 80-94 (1997).

16. Levitin, A., Redman, T.: Quality dimensions of a conceptual view. Information Processing & Management 31(1), 81-88 (1995).
17. Naumann, F., Rolker, C.: Assessment methods for information quality criteria. In: Proceedings of the 5th International Conference on Information Quality, Humboldt-Universität zu Berlin, Institut für Informatik, pp. 148-162 (2000).
18. OKF - Open Knowledge Foundation. Open Knowledge Definition. http://www.opendefinition.org/, last accessed 2017/9/17 (2006)
19. Omidyar Network: How Open Data Can Help Achieve the G20 Growth Target. www.omidyar.com/sites/default/files/file_archive/insights/ON/Report_061114_FNL.pdf, last accessed 2017/1/7 (2014).
20. Parks, W.: The open government principle: applying the right to know under the constitution. The George Washington Law Review 26(1), 1-22 (1957).
21. Pignotti, E., Corsar, D., Edwards, P.: Provenance Principles for Open Data. In: Proceedings of Digital Engagement 2011 (2011).
22. Scannapieco, M., Catarci, T.: Data quality under a computer science perspective. Archivi & Computer 2, 1–15 (2002).
23. Self, P.: Government by the Market? Macmillan, London (1993).
24. Stake, R.E. Qualitative case studies. In: Denzin, N.K., and Lincoln, Y. S. (Eds.), Strategies of qualitative inquiry, Sage, Los Angeles, pp. 119-149 (2008).
25. Tayi, G.K., Ballou, D.P.: Examining data quality. Comm.s of ACM, 41(2), 54-57 (1998).
26. TED: TED Processed Database: Notes & Codebook, Version 2.2. last accessed 2018/5/17: http://data.europa.eu/euodp/repository/ec/dg-grow/mapps/TED(csv)_data_information.doc
27. TED: TED Advanced Notes. Version 0.9. last accessed 2017/9/20: http://data.europa.eu/euodp/repository/ec/dg-grow/mapps/TED_advanced_notes.docx.
28. Tygel, A., Kirsch, R.: Contributions of Paulo Freire for a critical data literacy. In: Proceedings of Web Science 2015 Workshop on Data Literacy (2015).
29. van Zeist, R., Hendriks, P.: Specifying software quality with the extended ISO model. Software Quality Journal 5(4), 273-284 (1996).
30. Vetrò, A., Canova, L., Torchiano, M., Minotas, C.O., Iemma, R. Morando, F.: Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly 33(2), 325-337 (2016).
31. Wang, R.Y. Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12, 4, 5-33.
32. Wormell, I.: Information quality: definitions and dimensions. In: Proc.s of a NORDINFO Seminar, Royal School of Librarianship, Copenhagen, Taylor, London (1990).
33. Yin, R.K.: Case study research. Design and methods. Sage, Thousand Oaks, CA (2003).
34. Yu, H., Robinson, D.G.: The new ambiguity of open government. UCLA Law Review Discourse, 59, 178 (2011).
35. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. Quality assessment for linked data: A survey. Semantic Web 7(1), 63-93 (2016).
36. Zuiderwijk, A., Janssen, M.. Barriers and development directions for the publication and usage of open data: A socio-technical view. In: Open government, Springer, New York, pp. 115-135 (2014).
37. Zuiderwijk, A., Janssen, M.: The negative effects of open government data - investigating the dark side of open data. In: Proceedings of the 15th Annual International Conference on Digital Government Research, ACM, pp. 147-152 (2014).