DEVELOPING AN INDEX FOR MEASURING OGD PUBLISHER COMPLIANCE TO GOOD PRACTICE STANDARDS: INSIGHTS FROM OPENDATA.SWISS

Auriane Marmier & Tobias Mettler University of Lausanne, Swiss Graduate School of Public Administration, Switzerland

Citation: Marmier A, Mettler T (2020) Developing an index for measuring data quality of published open government data: Insights from opendata.swiss. Information Polity, 25(1), 91-110, DOI: 10.3233/IP-180120.

Published version at: https://content.iospress.com/articles/information-polity/ip180120

Abstract. In many countries, public organisations are among the largest creators and gatherers of data. To increase economic growth, governments have therefore begun to liberate access to large parts of government data by developing open government data (OGD) initiatives. Since the emergence of OGD initiatives, many OGD portals have been launched. There is a common belief that sharing OGD throughout platforms would be sufficient to motivate companies to re-use data and improve economic growth. However, there is very little evidence about the quality of shared OGD. For companies to be able to re-use, share and create value from OGD, data publishers must meet certain good practice standards. Following a pragmatic research approach, in this paper we present an index that can be applied for the quality assessment of the published OGD on portals. On the basis of 17,777 published data resources gathered from the Swiss OGD portal (opendata.swiss), we demonstrate the logic of the index and discuss the key learnings we obtained from applying the index to this concrete case. We conclude that, in Switzerland, the adherence to good practice standards for publishing OGD is fairly low.

Keywords. Good practice, liberation of data, open government data, standards.

1. INTRODUCTION

In the current economic context, in which data is driving innovation, many governments have perceived their data as a resource of strategic relevance (Bates, 2014; Munné, 2016). Through their daily activities, governmental authorities and public organisations generate and collect vast quantities of data (Pollock, 2011) and, over the past few years, governments have by far been the major creators of data (Máchová & Lnénicka, 2017). So that governmental authorities may benefit from such resources, they have begun to discuss the implementation of different initiatives. With the emergence of open movements, initiatives have been formed to liberate the access to large parts of government data (Attard, Orlandi, Scerri, & Auer, 2015; Kalampokis, Tambouris, & Tarabanis, 2011). These initiatives include open governmental data (OGD) platforms that allow everyone to have access to data were founded.

Torchiano, Vetrò, and Iuliano (2017) noted that many authors (C Alexopoulos, Loukis, & Charalabidis, 2014; Barry & Bannister, 2014; A Zuiderwijk, Janssen, van de Kaa, & Poulis, 2016) recommend facilitating access to public data sets. In their view, this could result in interesting new types of re-use for various (i.e. industrial, individual, scientist, etc.) actors. According to Davies (2010), accessible governmental data may generate important revenues, not only for the public but also for the private sector (e.g. applications, smart city tools). This data could allow the development of new value-added services, commercial purposes as well as political issues (Charalampos Alexopoulos, Zuiderwijk, Charapabidis, Loukis, & Janssen, 2014). Thus, the common belief that sharing OGD throughout platforms would be sufficient to motivate companies to re-use data and improve economic growth was born (Gascó-Hernández, Martin, Reggi, Pyo, & Luna-Reyes, 2018; Vickery, 2011). OGD platforms are digital infrastructures, which allow everyone to have access to the data, download them and use them for any purposes (Danneels, Viaene, & Van den Bergh, 2017; The European Commission, 2018). This belief has therefore led to major investment for the creation of many OGD platforms (Charalampos Alexopoulos et al., 2014; Jetzek, Avital, & Bjorn-Andersen, 2014; Ubaldi, 2013).

However, there is very little evidence that OGD platforms in fact enable data re-use and foster innovation and economic growth (Martin, 2014). Although several governments release a large amount of data open to the public, Danneels et al. (2017) ask about whether the OGD platforms will allow reaching the targets strategic goals and will live up to expectations. Many authors continue to express high hopes regarding the OD potential (Charalabidis et al., 2018; publishing, 2016; Anneke Zuiderwijk, Shinde, & Janssen, 2018a) but according to our knowledge, none of the studies has demonstrated the actual re-use of data available on OGD platforms (Danneels et al., 2017). According to van Veenstra and van den Broek (2013), it became imperative for governments to encourage the development of ways allowing practical OGD re-use as well as the reinforcement of platforms attractiveness. It is an illusion to believe that the publication of OD is automatically followed by the download and re-use of published data (Danneels et al., 2017).

Concerning non-re-use of OGD, one of the reasons explored in research relates to quality. According to Torchiano et al. (2017), public administrations (PAs) provide low-quality data. As Umbrich, Neumaier, and Polleres (2015) point out, low-quality of data can seriously affect their re-utilisation. Also, Allison (2010) stressed the fact that the quality of data on OGD platforms is not always appropriate for use in applications (e.g. non-machine-readable formats, no licence on the data sets). To address these quality issues, the literature has presented different models for analysing and evaluating OGD (Charalabidis et al., 2018; Conradie & Choenni, 2012; Heimstädt, Saunderson, & Heath, 2014). Researchers have analysed the many characteristics of OGD, have identified the dimensions that define their qualities, and have measured OGD quality. In the same vein, open data (OD) advocates have developed principles (e.g. use of commonly owned or open formats) encouraging organization to follow standards (e.g. type of formats such as .cvs, .txt. etc.), frequently inspired by data quality dimensions, and have developed indices in order to measure OGD platforms' quality. Although many papers and indices have measured data quality in terms of OGD platforms' usability (European Data Portal, 2017), availability (Open Data Monitor, 2018) or openness (Open Knowledge Fundation, 2018c),

few have focused on the quality of the data and none have reported on the data providers' compliance to these standards when publishing OGD.

To advance the research on the actual use of OGD platform, we propose and developed a *Compliance Index* **with the aim to analyse public organisations' compliance levels to existing OD standards**. In our view, the compliance to good practice standards on OGD platforms is essential to improve the quality of data publication, foster data exchange and consequently improve its re-use. We use the Swiss OGD platform *opendata.swiss* as an example and scrutinise the metadata published by the OGD platform, seeking to shed some light on this research question: *To what extent do data published by public organisations on OGD platforms comply to existing standards?*

The remainder of this paper is structured as follows. In Section 2, we briefly explore how data quality is treated in the literature and what dimensions are used most often in OD standards. Next, we describe the methodology used to develop our Compliance Index. We will then present the results of the index from its application in Switzerland, concluding with the discussion of our research's implications and limitations.

2. BACKGROUND

2.1. DATA QUALITY IN THE LITERATURE

In the literature, the definition of data quality is a recurring challenge. According to Sadiq, Yeganeh, and Indulska (2011), data quality has been studied largely over the two last decades (Wahyudi, Kuk, & Janssen, 2018; Zhang, Indulska, & Sadiq, 2019), without finding a clear consensus on the subject (Corsar & Edwards, 2017). As Sidi et al. (2012) noted, data quality depends on the context, as well as on the perspectives of data consumers, which complicates the process of defining quality. In the Data Management Book of Knowledge (edited by DAMA UK), data quality refers to "the characteristics associated with, and to the processes used to measure or improve the quality of data" (Askham et al., 2013). According to CKAN Association Steering Group (CKAN Organization, 2018), the data quality is a complex measure of data properties and utilises various dimensions. These

dimensions make it possible to know whether data are appropriate for their purpose. Strong, Lee, and Wang (1997) and Wang and Strong (1996) defined data quality in terms of their ability to be re-used by data consumers, while Wand and Wang (1996) defined data quality as high when the data meet their objectives. In the information system (IS) research domain, data quality is often presented as a multidimensional concept based on key dimensions (Abate, Diegert, & Allen, 1998; Fox, Levitin, & Redman, 1994; Huh, Keller, Redman, & Watkins, 1990; Redman & Blanton, 1997; Wand & Wang, 1996). IS researchers have also presented a wide range of dimensions that are useful in assessing this multidimensional concept of quality. For instance, some authors have mentioned dimensions such as accessibility, timeliness, completeness or amount of data, while others have focussed on data accuracy, consistency or timeliness (Wand & Wang, 1996). But as Batini, Cappiello, Francalanci, and Maurino (2009) noted, defining data quality and the dimensions chosen to assess it remains hard.

In this paper, we align ourselves with the general definition of data quality used in IS research. Our choice resonates with (1) the lack of consensus in the literature regarding data quality for the optimization of OGD platform' usage and (2) a pragmatic approach since the aim of this study is not to reflect on a consensus regarding data quality definitions. As a matter of fact, the literature shows that platforms - whether they are OD or OGD - do not use necessarily the same dimensions to measure data quality (H. C. Yang, Lin, & Yu, 2015). Máchová and Lnénicka (2017) found a lack of harmonisation on the dimensions used and underlined the need for quality standards. Furthermore, the literature shows that most of the studies focus on the evaluation of the platform itself rather than on the quality of its resources (Vetrò et al., 2016). To better understand which dimensions are used and at which level, we describe the different data quality dimensions proposed in the literature to assess OD and OGD platforms.

2.2. DATA QUALITY AND OPEN DATA

According to Z. Yang, Cai, Zhou, and Zhou (2005), data quality is one of the most relevant factors in the evolution of a web portal (Ciancarini, Poggi, & Russo, 2016). For several

authors, data quality is a serious risk for OD platforms; they argue that data quality could have impacts on OD projects' implementation (Open Knowledge Fundation, 2018b; Torchiano et al., 2017; Umbrich et al., 2015). Prior research (e.g. Henninger, 2013; Janssen, Charalabidis, & Zuiderwijk, 2012) pointed to the importance that published data (such as catalogues, dataset or resources) must be comprehensible, complete, consistent, and machine-readable. Also, an unclear licensing of the data or inconsistent pricing might additionally discourage the use of OGD (Vickery, 2011). With OGD initiatives, many OGD portals have been launched therefore, various efforts have been made to study the quality of open data. Scientists and data advocates have worked on many models that propose different dimensions for evaluating quality.

Umbrich et al. (2015) and Reiche and Höfig (2013) monitored and assessed the quality of platforms using quality dimensions similar to those in the IS literature. Umbrich et al. (2015) chose a total of six dimensions (data retrievability, use, completeness, accuracy, openness and contactability), while Reiche and Höfig (2013) focussed on quality dimensions such as formalism, completeness, accuracy, richness of information, accessibility and availability, among others. Both base their measurement on datasets but provide an evaluation of the quality of the platforms. Maurino, Spahiu, Batini, and Viscusi (2014) realised a study based on 50 OGD Italian datasets. They evaluate the quality in terms of completeness, accuracy and timeliness and also perform the quality evaluation at platforms level. Through their platforms analysis', only a few researchers propose solutions for the data providers. Reiche, Höfig, and Schieferdecker (2014) focussed on the assessment of OGD platforms' metadata to define their quality. Among others, they used dimensions such as the format used, the format's machine-readability, the licence, the URL's accessibility and the existence of a contact e-mail address. In order to give feedbacks to locals publishers, Neumaier, Umbrich, and Polleres (2016) proposed an instrument to evaluate open data portals in small and medium-size cities. Chatfield and Reddick (2017) did the same by providing a longitudinal cross-sector analysis for several cities in Australia. They used for these purposes dimensions such as policy intensity, open data provision and data format variety, among others.

Indices available on the Internet also used dimensions as metrics to measure some OGD platforms' quality (European Data Portal, 2017; Open Data Monitor, 2018), such as the data's completeness, machine-readability and licence. The Global Open Data Index (GODI) provides a snapshot of available OGD regarding different domains, such as government spending, draft legislation, election results, or land ownership (Open Knowledge Fundation, 2018c). The goal of GODI is to illustrate the range of data that is available on nation-wide OGD platforms. The Open Data Barometer aims at comparing the readiness, implementation, and impact of OGD platforms worldwide on an aggregate, generic level (World Wide Web Foundation, 2018). Again, the index does not deliver immediate information about the quality of shared data but relies on responses of intermediaries or operators of national OD platforms. Lastly, the Open Data Monitor shows indicators for measuring the readiness and maturity of OGD platforms across Europe (European Data Portal, 2017). Given that its goal is to compare the maturity level of European countries, it refrains from analysing data providers directly and rests on the level of national OGD platforms.

We noticed that, all of these studies and indices concentrate, first and foremost, on the *intermediaries or operators* of nation-wide OGD platforms and less on the process of data publication itself. To our knowledge, there is no studies or indices that focus on the compliance or adherence to good practice standards from the perspective of *data providers*. As discussed by Kubler, Robert, Neumaier, Umbrich, and Le Traon (2017), assuring a high and consistent data publication process (i.e. data quality, metadata access, standards applications, etc.) on OGD platforms is one of the main challenges today. Poor datasets or a lack of information may largely affect the discovery and processing of OGD; while some kind of data might be available on a nation-wide OGD platform, it **might not be used or found by machines and humans**. Zhang et al. (2019) underline that data users such as scientists and citizens should be able to explore, investigate and understand the quality of datasets to foster their re-use. We assume that even with a high degree of data quality, a low level of metadata requirements will negatively impact the use or reutilization of OGD. According to the W3C Egov Interest Group (W3C, 2017), without a proper

documentation of the nature and content of the data it is hardly reusable. To guarantee a more effective use of OGD platforms, we strongly believe that good descriptions of metadata are crucial: they are the keys that allow data consumers to explore and understand the signification of data and evaluate if the provided quality is suitable for their purposes. We presume that ensuring data quality begins by ensuring that good practices standards are respected and applied by data providers, but also ensuring an access to a comprehensive documentation.

2.3. DATA QUALITY AND STANDARDS

Data are described as the most valuable asset of the century. Nevertheless, many authors are raising questions about why the promised data-driven innovation still not happening. One of the hypotheses is that data quality dimension do not follow standards and that could impact data exchange (W3C, 2017). The Web Best Practices Working Group studied 26 OD use cases in order to understand how the lack of standards could retard the development of the data-driven economy (W3C, 2017). Umbrich et al. (2015) compared 82 OGD platforms' and showed the utilisation of a wide range of formats. He explains this diversity by a lack of standards defining the resources formats. Many authors further notice claim the necessity to standardize the dimensions used to assess data quality (Ciancarini et al., 2016; Máchová & Lnénicka, 2017; Vetrò et al., 2016; H. C. Yang et al., 2015). Thus, in our view, governments' application of and compliance to good practice standards on OGD platforms are essential to improving the platforms' quality as well as to attaining OGD objectives (e.g. re-use of their content).

As seen above, although there is still no consensus on how to define data quality for OGD portals, and even less on the dimensions used to measure data quality, we saw that several data quality dimensions have emerged from the OGD literature. Given that there is an unstructured development of data quality literature, Paré, Trudel, Jaana, and Kitsiou (2015) recommend conducting a comprehensive narrative literature review to manually gathered dimension that occurred the most. Therefore, we listed the different dimensions of data quality used by researchers to assess OGD platforms' characteristics in Table 1. Among

them, several form part of the good practices published by advocates of OD and the most frequently used dimensions to measure data quality in the OGD context are: licence, easy access, machine-readable format, timeliness and completeness.

DATA QUALITY DIMENSIONS	(Vetrò et al., 2016)	(Máchová & Lnénicka, 2017)	(Reiche & Höfig, 2013)	(Umbrich et al., 2015)	(Reiche et al., 2014)	(Sunlight Foundation, 2018)	(Open Knowledge Fundation, 2018c)	(Open Data Monitor, 2018)	(European Data Portal, 2017)	(World Wide Web Foundation, 2018)
ACCESSIBLE			x							x
ACCURACY	x		x	x						
AVAILABILITY		x	x					x	x	
COMPLETENESS	x		x	x		x				x
COMPLIANCE	x									
CONSISTENCY	x									
DATASET URL		x			x					
EASY ACCESS				x	x	x	x		x	x
FREE USAGE COST						x	x		x	x
LICENSE				x	x	x	x	x	x	x
MACHINE READABLE FORMAT					x	x	x	x	x	x
METADATA AVAILABILITY			x							
METADATA COMPLETENESS			x					x		
NON-DISCRIMINATION						x				
OPEN FORMAT						x	x		x	x
OPENNESS				x						
PERMANENCE						x				
PRIMARY/ PRIMACY						x				
RE-USABILITY OF DATA			x	x					x	
TIMELINESS/ UP TO DATE	x	x				x	x		x	x
USE OF DATA				x					x	

Table 1: Most frequently used dimensions to measure data quality in OGD

By listing dimensions that occur the most in the literature, we rapidly realized that most of the authors use different dimensions to describe the same idea. For instance, we constantly encounter the dimensions *availability, accessibility or re-usability of the data, which* without clear definition may lead to the same interpretation. This is maybe the reason why the ODI (Open Data Index, 2016) recommended that researchers, developers and policy-makers adhere to common data standards. In light of the foregoing considerations, we chose to follow the good practice standards of the Sunlight Foundation (SF), testing whether they were respected by Swiss OGD platform publishers. Its principles have been developed in order to "*empower the public's re-use of public data held by governments*". They re-group most of the important open source publishing rules' (Sunlight Foundation, 2018) and cover a great variety of dimensions such as data accessibility, availability, technicality and legality. Even if the vocabulary used is not always similar, we also chose to follow the OGD good practice standards propagated by the SF (Sunlight Foundation, 2018) because its principles correspond to the quality dimensions most often cited in the literature and most often recommended by the most advanced OGD platforms (e.g. *data.go.uk, data.gouv.fr, data.gov, etc.*). Contrarily to other recommendations, the meaning of each 10 principles is deeply and comprehensively defined. While some principles meaning could be discussed (e.g. why speaking about usage cost when we speak about open data, or commonly owned formats such as Excel while it is subject to licence requirements), these 10 principles provide a very good representation of the challenges of publishing and using OGD, from the perspectives of both data providers and consumers. In our view, they remain essentials to ensure feasibility and continuity in the OD re-use. Finally, the SF is one of the first foundations to take an active interest on the special case of government data. We summarise the 10 principles in Table 2 (for a full-length description of the principles, see Appendix A).

#1 Completeness	Resources published on OD platforms should contain all raw information and metadata defining and explaining their content.
#2 Primacy	Resources published on OD platforms should also include the original information released by the government.
#3 Timeliness	Resources should be available to the public in a timely manner.
#4 Easy access	Resources published on OD platforms should be easy to find and download.
#5 Machine-readable format	Resources should be stored in a machine-readable format (i.e. should be processable by a computer)
#6 Non-discrimination	Resources published on OD platforms should be accessible without having to identify oneself (e.g. via needing to log in) or having to provide a justificatory reason.
#7 Open format	Resources should be usable without proprietary software.
#8 Open licencing	Resources published on OD platforms should use an open licencing model.
#9 Permanence	Resources published on OD platforms should be accessible by machines and humans over time.
#10 Usage cost	Resources should be available for free.

Table 2: Summary of the 10 Sunlight Foundation's principles for data quality

Usually, only open data portals on the national level are evaluated and very little is known from portals at the regional, local or organizational level. The approach we present, which we named the Compliance Index, analyses whether Swiss public organisations are complying with the 10 principles recommended by the SF. We proposed that index in order to go further in data quality measurement and also to fill the gap in data quality evaluation at the organizational level. We chose to apply the Compliance Index to Switzerland, partly because the Swiss OGD platform *opendata.swiss* have been launched in 2016 and only two years after its opening, authorities already observe trouble with its use. Furthermore, the first Swiss OGD strategy (from 2014 to 2018) is coming to an end, and that is, therefore, an appropriate time to evaluate it but also to propose an appropriate recommendation to the authorities in charge of the platform and data providers. Due to its federal structure, Switzerland offers to our study different analysis levels allowing us to observe the OGD publication through three types of governance (i.e. federal, cantonal and municipal).

3. METHODOLOGY

In developing an index that can be used for the quality assessment of data publication, we chose to follow a pragmatic epistemology (Goldkuhl, 2012). The essence of pragmatic research lies in the interplay between actions and change: To alter certain aspects of reality (in our case data quality of published OGD), actions are required. Empirical evidence is essential to initiate change and possibly alter the current situation to a more desired state. Actions and their impact can also contribute to further cognitive clarification and development. This way of researching contrasts with, for example, purely descriptive research that primarily seeks to explain reality by using models (or a structure of relations) and which uses methods that emphasize the discovery of new knowledge and verify existing (structural) knowledge without deliberately distorting reality. In this sense, pragmatic research is problem-driven research by necessity since complex problems do not respect philosophical, historical, or disciplinary boundaries of science. That said, we followed the common steps of pragmatic research, which can be summarized as problem analysis, artefact construction, artefact application, and interpretation and learning from the results. After having outlined the current issues regarding data quality and OGD in the previous sections, we now turn to the description of the construction of our compliance index and show its application based on the Swiss context.

3.1. INPUT DATA FOR BUILDING THE COMPLIANCE INDEX

Data collection. To see whether governmental authorities and Swiss public administrations are following the aforementioned OGD good practice standards, we analysed the available metadata that is published alongside a data resource and that usually can be downloaded from an OGD platform. Like many OGD platforms worldwide (Kirstein et al., 2019), *opendata.swiss* uses the Comprehensive Knowledge Archive Network (CKAN), a web-based open-source management system to store and distribute OGD (Open Knowledge Fundation, 2018a). We harvested the available CKAN metadata, such as resources titles, resources authors name or resources formats, by creating specific requests so as to obtain a valid catalogue end-point for extracting all resources-related metadata. All open data catalogues generated with CKAN are organised by group, datasets and resources. One group contains many datasets and one dataset may contain at least one or more resources. We first used the application programming interface (API) offered by CKAN to create requests and then used the excel tool “power pivot” to download the metadata of all

published resources. The data downloading was done on 26 May 2018, when the portal contained entries from 44 public organisations – a total of 17,777 data resources.

Data manipulation and imputation of missing metadata values. In phase 2, we extracted and analysed all relevant resources by means of Excel using the power pivot functionality. To remove inconsistencies in the obtained data, we cleaned and pre-selected specific CKAN metadata fields in Excel on the basis of their accessibility, interpretability and coherence in terms of the 10 principles. The final CKAN metadata fields we used to create the Compliance Index appear in Appendix B. In case that metadata had missing values for some resources, we chose the single imputation method, including the hot deck imputation solution. This approach considers missing values as part of the analysis by imputing them another value. Thus, in case of a missing value, we imputed a negative answer (*no*), as we will now explain.

3.2. TREATMENT OF THE METADATA

Questions set. The coding of the metadata was realised via a set of questions we developed on the basis of the 10 principles (for more details, see Appendix C). For instance, to assess the compliance to Principle 5 (machine-readability), we asked whether the resources were available in a format such as RTF, CSV, XML or JSON, i.e. whether the value of the CKAN metadata field `resources.format` corresponded to one of the pre-defined machine-readable formats. Such questions examined whether the metadata (and metadata value) adhered to the good practice standards of publishing OGD. In this sense, each of the 10 principles is linked to a specific question. A question can engage one or many CKAN metadata fields if it is not mutually exclusive. For instance, the CKAN metadata field `resources.rights` was used in multiple questions to answer the adherence to Principles 8 (open licencing) and 10 (usage costs).

Allocations of points. As noted, we transformed the answers into binary numbers. A positive answer (*yes*) concerning the application of a principle to a resource = one point, while a negative answer (*no*) = zero points. By using binary numbering, nominal answers

easily translated in ordinal answers. This ensured a computation of the scores obtained for each resource (index per resources) and the creation of an index for organisations (Compliance Index). To compute the index per resources and the Compliance Index, we opted for the equal weighting method. This method implies an equal status for all the principles. In this sense, the index per resources and the final Compliance Index considered all the principles as equally important concerning publishing OGD. Thus, these two indices do not favour certain principles over others.

Scoring of the Compliance Index. For each published resource by an organisation, every principle grants a number of points; the sum of the points constitutes the score of the index per resources. On this basis, we computed the Compliance Index per organisation, which corresponds to the average index scores per resources. Considering our study purpose, a linear aggregation method allowed for the computation of an index per resource, but also provides information on principles applied to a given resource. For instance, a score of seven on the index per resources means that a given resource applied seven of the 10 good practice standards. This also indicates that the organisation that publishes these resources must modify some of its practices if it is to be considered a fully compliant OGD provider. In sum, the Compliance Index provides a general overview of how organisations publish OGD, and the index per resources allows for a deeper examination of how each organisation manages its resources.

4. RESULTS

We will now present the results of applying our method to data providers publishing their data on the Swiss OGD platform *opendata.swiss*. Figure 1 provides a general overview of the current compliance levels regarding all 10 principles over all 17,777 data resources. Segments in blue represent the percentage of resources that respected a specific principle, while segments in yellow constitute the percentage of resources that did not. As the illustration suggests, not all good practice standards were followed equally. On *opendata.swiss*, data resources only respected the principles of primacy and licencing to 100%. Only two other principles, completeness and open standards, were respected by

more than 50% of data resources. Permanence, timeliness and non-discrimination were given in only less than 10% of the data resources.

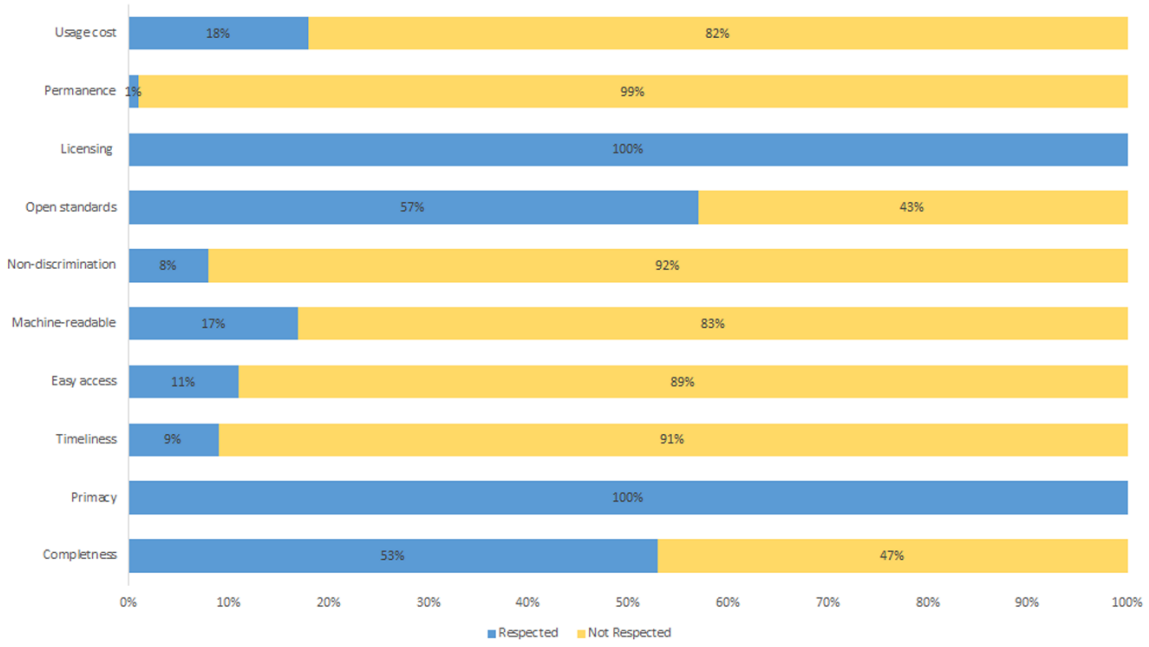
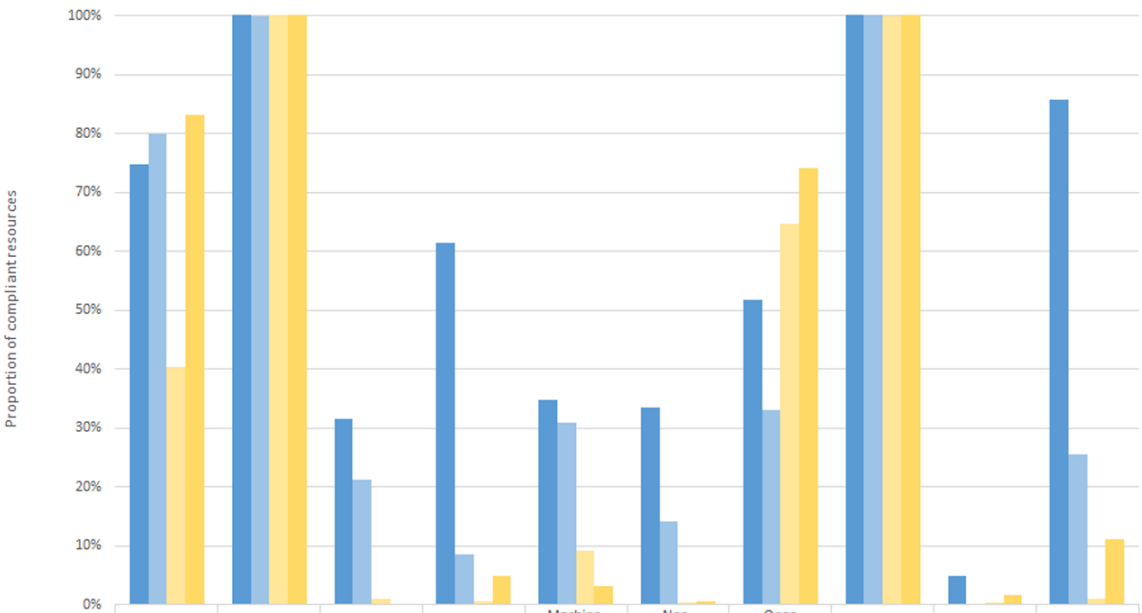


Figure 1: General overview of adherence to the 10 good practice standards

Figure 2 shows the percentage of data resources that complied to the 10 principles by the data provider's political level (communal, cantonal, federal and other). As we can see, communes tended to follow good practice standards more than the three others – it was the only group of data providers that largely applied the easy access and usage cost principles. At the canton level, although some principles were not respected as much as on the communal level, we could still determine efforts to comply with 9 out of 10 principles. At the federal level, only the principles of completeness, primacy, open standards and licencing were followed.

Completeness, primacy, machine-readability, open standards and licencing are the principles that were most followed by data providers across most political levels. For instance, completeness was largely followed by communal (75%), cantonal (80%) and

other public actors (83%), while these percentages only reached 35%, 31% and 3% in the case of machine-readability.



	Completness	Primacy	Timeliness	Easy access	Machine-readable	Non-discrimination	Open standards	Licensing	Permanence	Usage cost
commune	75%	100%	32%	61%	35%	33%	52%	100%	5%	86%
canton	80%	100%	21%	9%	31%	14%	33%	100%	0%	26%
confederation	40%	100%	1%	1%	9%	0%	65%	100%	0%	1%
other	83%	100%	0%	5%	3%	1%	74%	100%	2%	11%

Figure 2: Compliance of data resources by political level

Table 2 shows the percentages of data resources that were respected for each principle based on OGD categories (e.g. data for tourism, trade, health-care or agriculture). We did this analysis in order to understand possible differences in publishing OGD across different public sector domains, assuming that data providers face different complexity levels (e.g. different regulations concerning privacy and data protection).

Despite not finding a clear pattern from this analysis, some OGD categories stood out. For instance, a small share of administrative data resources seemed to respect all good practice standards and may, overall, be more compliant than other categories. To a lesser extent than administrative data resources, OGD categories such as geography, legislation and territory also adhered to most of the principles, while educational, energy and industrial data resources appeared to have more difficulties applying the principles. Surprisingly, a

principle may be broadly respected in one domain, yet may be completely neglected in others. For instance, the principle of completeness was applied in 92% of data resources published in the legislation category, while only 17% of the data resources under the trade category provided indications about their completeness. Another example is the principle relating to data resources' usage costs. While 67% of data resources published under the trade category respected this principle, the percentage only reached 2% in case of data resources in the industrial category.

Principles: / OGD Categories:	Completness	Primacy	Timeliness	Easy access	Machine-readable	Non-discrimination	Open standards	Licensing	Permanence	Usage cost
administration	83%	100%	4%	34%	36%	23%	53%	100%	13%	50%
agriculture	50%	100%	3%	5%	11%	4%	53%	100%	1%	6%
construction	44%	100%	8%	14%	12%	5%	47%	100%	5%	18%
crime	59%	100%	0%	0%	17%	0%	68%	100%	1%	3%
culture	64%	100%	1%	5%	14%	3%	64%	100%	4%	11%
education	27%	100%	1%	3%	6%	2%	64%	100%	1%	6%
energy	60%	99%	**0%**	3%	8%	2%	21%	100%	0%	8%
finances	83%	100%	**54%**	6%	4%	2%	20%	100%	7%	13%
geography	61%	100%	2%	22%	33%	27%	43%	100%	0%	49%
health	54%	100%	6%	4%	6%	2%	80%	100%	0%	6%
industry	21%	100%	1%	0%	5%	0%	55%	100%	1%	**2%**
legislation	**92%**	100%	0%	8%	32%	0%	56%	100%	5%	15%
mobility	68%	100%	3%	25%	10%	8%	50%	100%	0%	30%
national economy	57%	100%	1%	14%	13%	3%	66%	100%	2%	16%
politics	84%	100%	4%	5%	39%	2%	50%	100%	3%	8%
population	34%	100%	3%	6%	9%	1%	60%	100%	0%	7%
prices	48%	100%	2%	3%	10%	2%	79%	100%	0%	9%
public order	91%	100%	0%	1%	34%	1%	54%	100%	3%	8%
social security	48%	100%	0%	10%	14%	7%	62%	100%	0%	16%
statistical basis	41%	100%	1%	1%	2%	1%	91%	100%	0%	2%
territory	71%	100%	1%	17%	28%	12%	48%	100%	0%	24%
tourism	31%	100%	0%	3%	6%	3%	73%	100%	0%	10%
trade	**17%**	100%	0%	0%	0%	0%	58%	100%	17%	**67%**
work	55%	100%	3%	0%	17%	0%	68%	100%	0%	2%

Table 2: Compliance to data resources by OGD categories

Table 3 provides an overview of the final scores of the Compliance Index per public organisation and political levels. For anonymisation reasons, organisation names are not displayed and have been replaced by an identification number. To recap, the Compliance Index is an average of the scores obtained by every resource (published by an organisation) for each principle. For the analysed timeframe, we found that the maximum number of principles respected by organisations was 5.9 and that the minimum number

was 2. Only 16% of public organisations complied with five or more good practice standards. Our results also indicate that the maximum principles respected by a data provider at the canton level was almost 6 and that the minimum was 2. Concerning the other political levels, a minimum of 2 principles were respected, but the maximum was only 4.7 principles, with a mean index score of 3.5. On average, data providers at the communal level were more compliant with the 10 principles (4.2). Data providers at the canton and federal levels respected on average 3.8 principles. A lower mean index score was only reached by public organisations not attributable to one of the aforementioned political levels (3.5). Indicatively, the average score of all resources published on the *opendata.swiss* platform, independently of their political level, was 3.75 out of 10. In concrete terms, this means that public organisations in Switzerland only respected about 4 out of 10 principles when publishing OGD.

ID	Federal organisations
1	**2.5**
2	2.8
3	2.9
4	2.9
5	3.1
6	3.1
7	3.2
8	3.2
9	3.2
10	3.3
11	3.3
12	3.4
13	3.5
14	3.9
15	4.0
16	4.5
17	4.6
18	5.0
19	5.6
20	5.6
21	**5.6**

Average	
	3.8

ID	Cantonal organisations
22	**2.0**
23	2.5
24	3.1
25	3.5
26	3.5
27	4.0
28	4.0
29	4.0
30	5.8
31	**5.9**

Average	
	3.8

ID	Communal organisations
32	**2.5**
33	4.0
34	4.0
35	4.0
36	4.2
37	4.8
38	**5.9**

Average	
	4.2

ID	Other organinsations
39	**2.0**
40	3.0
41	3.3
42	3.8
43	4.2
44	**4.7**

Average	
	3.5

Table 3: The Compliance Index scores for every organisation

5. DISCUSSION AND CONCLUSION

Our analysis results revealed that, in Switzerland, the adherence to good practice standards for publishing OGD is fairly low. Our Compliance Index also demonstrated that not all principles were equally well implemented by data providers. For instance, while we found that all resources published on *opendata.swiss* followed the primacy principle as well as indicated a licencing model, the principle of permanence saw less respect.

A possible interpretation for this disparity could be the fact that data providers allocate different importance to each principle. For instance, disclosing the correct licencing model is essential in order to inform data consumers about their data re-use options or could also be motivated by self-interest to protect current or future copyright interests. Conversely, respecting the permanence principle means keeping data resources available online for data consumers over an indefinite period. This is complex and costly for data providers, and requires a good understanding of existing data storage and network structures.

Further, our coding procedure could have led to this impression. Not all principles were always equally measured. For some principles, we needed to develop approximations based on multiple metadata fields, since no single field explained the principle in its entirety. For others (e.g. open format) only one metadata field was sufficient. Sometimes, the range of allowed values in determining compliance was unambiguous. For instance, the principle of timeliness can be interpreted in different ways (Emran, 2015). Vetrò et al. (2016) defined timeliness as the presence or absence of the updated version of a data set, while Atz (2014) measured it via the percentage of a data sets' up-to-dateness. One other case, machine-readable formats, is considered to be common knowledge, yet there is no standard coding list we could fall back on.

But there could also be a much simpler explanation. It could just be that the platform intermediary (i.e. the Swiss Federal Archives) marked certain metadata fields as compulsory, while making others optional for data providers. Accordingly, it would be impossible for data providers to publish OGD without for instance determining a licencing model or indicating the resource's originator. Since the existing technical and

organisational guidelines (e.g. DCAT-AP for Switzerland) or end-user handbooks are not always written in ways that can be understood by a non-technical expert in a public administration, it could be that the technical complexity of publishing OGD, paired with ignorance, may have influenced data providers to only publish metadata they could fully understand.

When looking at the differences at the distinct political levels, we see an interesting result. On average, data resources published by communal organisations are more compliant, particularly in terms of ease of access, non-discrimination and usage costs. A plausible explanation for this could be that these actors are closer to citizens and are therefore more attentive to their needs, or that federal agencies must deal with a greater diversity and complexity of data resources, making publishing OGD more difficult and error-prone. However, it could also be that certain actors (e.g. Federal Office of Statistics, Federal Office of Topography) are not only participating in sharing OGD because they want to (e.g. as communal or cantonal organisations), but because they have to. Enforcing top-down diffusion of OGD could strengthen the belief that OGD is a burden to public administrations, rather than a service to citizens (Janssen et al., 2012). Nonetheless, our results are surprising, given that prior research has found that federal-level governmental agencies often have a higher maturity and readiness to diffuse OGD than communal or regional (cantonal) public administrations (Anneke Zuiderwijk, Shinde, & Janssen, 2018b).

We have developed a Compliance Index and have conducted a comprehensive empirical analysis on the basis of metadata published on *opendata.swiss*. The purpose was to see whether public organisations in Switzerland followed good practice standards when they publish their data as open access (and not to determine what defines data quality in the OGD context). We developed the Compliance Index mostly because we observed that OGD re-use might increase only if data providers improve the quality of the data, and consequently applied good practice standards. We argue that data providers such as public organizations, cities and municipalities, among others, not only need quality platforms analysis but also clears and details feedbacks on their resources. For this purpose,

organizations need to know which resources in which dataset does not have an adequate format or is not up-to-date. This is one objective of the Compliance Index. In our view, our Compliance Index is a promising first step towards better understanding data quality dimensions and sharing practices of governmental authorities and public organisations on OGD platforms. It offers the possibility for data providers to clearly and rapidly identify their deficiencies or potential *OGD champions* who have a better understanding of the OGD publishing process. Further, combining with a platform traffic analysis the Compliance Index can also be used as a first approximation to understand what quality dimensions are the most useful to allow data to be downloadable and potentially re-used and, thus, may be of particular interest to data consumers and platform operators. Although several authors explored the question on the data quality, most of the evaluations investigated on open data platform's quality at national level (Máchová & Lnénicka, 2017; Open Knowledge Fundation, 2018c; Ubaldi, 2013) and only a few studies examined metadata resources quality's at the organization level (Berners-Lee, 2006; Maurino et al., 2014). Our Compliance Index draws on the studies of Vetrò et al. (2016) and Maurino et al. (2014) but propose to go deeper by analysing automatically the OGD platform content, from data providers perspective. It directly evaluated the published CKAN metadata of data resources to then performed evaluation at dataset and resources levels. An analysis on this level of granularity not only supports distinct data providers to improve certain aspects of their OGD publishing procedures, but also, intermediaries and operators of OGD platforms by receiving key performance indicators that helps them to monitor and understand their platform's current state of OGD quality. From the perspective of public managers, such an index could support planification and process organization. As a matter of fact, public managers could gain better knowledge on the evolution of the "opening process" their services and be able to identify the next steps for further implementing open practices. The compliance index allows a dual action approach composed of identifying problems regarding data publication and resolution. The compliance index could also reinforce trust regarding the actions of governments and politicians. For instance, thanks to this tool, citizen could see the evolution of the open government strategy and the

corresponding level of realization in practice. By making the compliance index available to OGD organizations, public managers could better collaborate with each other and therefore, strengthen the most effective techniques that promote data publication (e.g. anonymization rules, recurring problems, etc.). Furthermore, CKAN is currently the most used web application in the public sector, such as used in the United Kingdom, United States, and many other countries (W3C, 2019), for building open data catalogues (Kirstein et al., 2019). Thus, developing an index based on CKAN allows to not only improve OGD re-use in Switzerland but also other for other countries, with the aim to emphasize the quality assessment of data publication.

Since developing the Compliance Index required compromises and approximations, our study has limitations. First, we narrowed the analysis to the CKAN metadata available on *opendata.swiss*, because this was easier to extract and scrutinise than analysing the published data sets, which would have required multiple different approaches, including text mining, image mining and others, given that a wide variety of administrative, geographical and statistical data sets are made available by different public organisations. Also, determining content's quality is often subjective and highly context-dependent. In this sense, our Compliance Index remains a proxy to measure whether or not the dimensions that define data quality are well respected by governmental authorities and public organisations, but is nonetheless based on the analysis of actually published data and not on survey data about attitudes, perceptions or 'wishful thinking' of interested or involved (and thus not always unbiased) respondents. Second, we based our understanding of good practice wholly on the 10 principles defined by the Sunlight Foundation. We acknowledge that there are different standards and guidelines, suggesting fewer or more circumstantial principles for directing OGD initiatives. However, the SF regroups 30 open government advocates coming from fields such as research, public administrations and internet. We are convinced that they appear a good starting point. Third, there were also issues concerning operationalisation of measurements and the coding procedure. Some of the principles, such the ease of physical and electronic access, could not be examined in much depth, given that user feedback is not collected

systematically on *opendata.swiss* (we simply assumed that it is easier and more practical for end-users to download data directly from the platform than to be re-directed to another download website). Fourth, we listed the most used OD quality dimensions by realising a comprehensive narrative literature review. However, a systematic literature review would be more appropriate and could provide a more accurate idea of the different dimensions used. Then, we used binary operators when deciding whether or not a principle was respected. In certain situations, this may be a bit simplistic, since there are different levels of adherence or non-adherence (e.g. free re-use of data after registration is still better than paid use of OGD only). Finally, the use of the same metadata fields to approximate principles could cause multicollinearity problems. Using the same fields to define different principles may provoke a chain reaction, if one organization score low in one aspect. For example, organizations with empty fields in a certain category (e.g. `resources.rights`) will be penalized as often as this field is used.

A more nuanced view, particularly in this regard, would be helpful. We trust that our paper will motivate other researchers to delve into the complex dynamics and controversial nature of OGD quality and of ways to measure it. It is only if we have a proper approximation to measure OGD quality that it is possible to make reasonable projections on OGD's value and to corroborate the currently uncontested relationships between OGD, innovation and economic growth.

REFERENCES

Abate, M., Diegert, K., & Allen, H. (1998). A hierarchical approach to improving data quality. *Data Quality, 4*(1), 365-369.

Alexopoulos, C., Loukis, E., & Charalabidis, Y. (2014). A platform for closing the open data feedback loop based on Web2. 0 functionality. *JeDEM-eJournal of eDemocracy and Open Government, 6*(1), 62-68.

Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). *Designing a second generation of open data platforms: Integrating open data and social media.* Paper presented at the International Conference on Electronic Government, Berlin, Heidelberg.

Allison, B. (2010). My data can't tell you that. *Open government-Collaboration, transparency, and participation in practice*, 257-265.

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., . . . Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. *DAMA UK Working Group*, 432-435.

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly, 32*(4), 399-418.

Atz, U. (2014). *The tau of data: A new metric to assess the timeliness of data in catalogues*. Paper presented at the Conference for E-Democracy and Open Governement.

Barry, E., & Bannister, F. (2014). Barriers to open data release: A view from the top. *Information Polity, 19*(1, 2), 129-152.

Bates, J. (2014). The strategic importance of information policy for the contemporary neoliberal state: The case of open government data in the United Kingdom. *Government Information Quarterly, 31*(3), 388-395.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR), 41*(3), 1-16.

Berners-Lee, T. (2006). *Linked data-design Issues* Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). Open data evaluation models: Theory and practice. In *The World of Open Data* (pp. 137-172): Springer International Publishing.

Chatfield, A. T., & Reddick, C. G. (2017). A longitudinal cross-sector analysis of open data portal service capability: The case of Australian local governments. *Government Information Quarterly, 34*(2), 231-243.

Ciancarini, P., Poggi, F., & Russo, D. (2016). *Big data quality: A roadmap for open data*. Paper presented at the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService).

CKAN Organization. (2018). Data quality, what is it? Retrieved from https://ckan.org/2011/01/20/data-quality-what-is-it/

Conradie, P., & Choenni, S. (2012). *Exploring process barriers to release public sector information in local government*. Paper presented at the Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance.

Corsar, D., & Edwards, P. (2017). Challenges of open data quality: more than just license, format, and customer support. *ACM Journal of Data Information Quality, 9*, 1-3.

Danneels, L., Viaene, S., & Van den Bergh, J. (2017). Open data platforms: discussing alternative knowledge epistemologies. *Government Information Quarterly, 34*(3), 365-378.

Davies, T. (2010). *Open data, democracy and public sector reform: a look at open government data use from data.gov.uk*. Oxford, UK: University of Oxford.

Emran, N. A. (2015). Data completeness measures. In *Pattern Analysis, Intelligent Security and the Internet of Things* (pp. 117-130): Springer.

European Data Portal. (2017). Open data in Europe. Retrieved from https://www.europeandataportal.eu/en/dashboard#2017

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information processing & management, 30*(1), 9-19.

Gascó-Hernández, M., Martin, E. G., Reggi, L., Pyo, S., & Luna-Reyes, L. F. (2018). Promoting the use of open government data: Cases of training and engagement. *Government Information Quarterly, 35*(2), 233-242.

Goldkuhl, G. (2012). Pragmatism vs Interpretivism in Qualitative Information Systems Research. *European Journal of Information Systems, 21*(2), 135-146.

Heimstädt, M., Saunderson, F., & Heath, T. (2014). *Conceptualizing open data ecosystems: A timeline analysis of open data development in the UK*. Paper presented at the Conference for E-Democracy and Open Governement.

Henninger, M. (2013). The value and challenges of public sector information. *Cosmopolitan Civil Societies: An Interdisciplinary Journal, 5*(3), 75-95.

Huh, Y., Keller, F., Redman, T. C., & Watkins, A. (1990). Data quality. *Information and software technology, 32*(8), 559-565.

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

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2014). Data-driven innovation through open government data. *Journal of theoretical and applied electronic commerce research, 9*(2), 100-120.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). *Open government data: A stage model*. Paper presented at the International Conference on Electronic Government.

Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., & Hauswirth, M. (2019). *Linked Data in the European Data Portal: A Comprehensive Platform for*

Applying DCAT-AP. Paper presented at the International Conference on Electronic Government, San Benedetto Del Tronto, Italy.

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2017). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly, 35*(1), 13-29.

Máchová, R., & Lnénicka, M. (2017). Evaluating the quality of open data portals on the national level. *Journal of theoretical and applied electronic commerce research, 12*(1), 21-41.

Martin, C. (2014). Barriers to the open government data agenda: Taking a multi-level perspective. *Policy & Internet, 6*(3), 217-240. doi:10.1002/1944-2866.POI367

Maurino, A., Spahiu, B., Batini, C., & Viscusi, G. (2014). *Compliance with open government data policies: an empirical evaluation of italian local public administrations*. Paper presented at the Twenty Second European Conference on Information Systems.

Munné, R. (2016). Big data in the public sector. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe* (pp. 195-208): Springer.

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. *Journal of Data Information Quality, 8*(1), 2.

Open Data Index. (2016). Benchmarking data automatically. Retrieved from http://oldsite.theodi.org/guides/benchmarking-data-automatically

Open Data Monitor. (2018). The open data monitor. Retrieved from https://opendatamonitor.eu/frontend/web/index.php?r=dashboard

Open Knowledge Fundation. (2018a). Comprehensive knowledge archive network. Retrieved from https://github.com/ckan/ckan

Open Knowledge Fundation. (2018b). The Open definition. Retrieved from http://opendefinition.org/

Open Knowledge Fundation. (2018c). Tracking the state of open government data. Retrieved from https://index.okfn.org/place/

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information Management, 52*(2), 183-199.

Pollock, R. (2011). Building the (open) data ecosystem. Retrieved from https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/

publishing, O. (2016). Economic ans social benefits of internet openness. *OECD Digital Economy Paper, no 257*.

Redman, T. C., & Blanton, A. (1997). *Data quality for the information age*: Artech House, Inc.

Reiche, K. J., & Höfig, E. (2013). *Implementation of metadata quality metrics and application on public government data*. Paper presented at the Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual.

Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). *Assessment and visualization of metadata quality for open government data*. Paper presented at the Conference for E-Democracy and Open Governement.

Sadiq, S., Yeganeh, N. K., & Indulska, M. (2011). *20 years of data quality research: themes, trends and synergies*. Paper presented at the Proceedings of the Twenty-Second Australasian Database Conference-Volume 115, Perth, Australia.

Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). *Data quality: A survey of data quality dimensions*. Paper presented at the Information Retrieval & Knowledge Management (CAMP), 2012 International Conference.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Sunlight Foundation. (2018). Ten principles for opening up government information. Retrieved from https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

The European Commission. (2018). Open Data portals. Retrieved from https://ec.europa.eu/digital-single-market/en/open-data-portals

Torchiano, M., Vetrò, A., & Iuliano, F. (2017). *Preserving the benefits of open government data by measuring and improving their quality: An empirical study*. Paper presented at the Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual.

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance*(22), 0-1.

Umbrich, J., Neumaier, S., & Polleres, A. (2015). *Quality assessment and evolution of open data portals*. Paper presented at the Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference.

van Veenstra, A. F., & van den Broek, T. A. (2013). *Opening moves–drivers, enablers and barriers of open data in a semi-public organization*. Paper presented at the International Conference on Electronic Government, Berlin, Heidelberg.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly, 33*(2), 325-337.

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

W3C. (2017). Data on the Web Best Practices. Retrieved from https://www.w3.org/TR/dwbp/

W3C. (2019). Comparing DKAN and CKAN. Retrieved from https://docs.getdkan.com/en/latest/introduction/dkan-ckan.html

Wahyudi, A., Kuk, G., & Janssen, M. (2018). A process pattern model for tackling and improving big data quality. *Information Systems Frontiers, 20*(3), 457-469.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

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.

World Wide Web Foundation. (2018). Open data barometer. Retrieved from https://opendatabarometer.org/?_year=2016&indicator=ODB

Yang, H. C., Lin, C. S., & Yu, P. H. (2015). *Toward automatic assessment of the categorization structure of open data portals*. Paper presented at the International Conference on Multidisciplinary Social Networks Research, Berlin, Heidelberg.

Yang, Z., Cai, S., Zhou, Z., & Zhou, N. (2005). Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information & management, 42*(4), 575-589.

Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems. *J Business Information Systems Engineering, 61*(5), 575-593.

Zuiderwijk, A., Janssen, M., van de Kaa, G., & Poulis, K. (2016). The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments. *Information Polity, 21*(3), 223-236.

Zuiderwijk, A., Shinde, R., & Janssen, M. (2018a). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? *International Review of Administrative Sciences*, 0020852317739115.

Zuiderwijk, A., Shinde, R., & Janssen, M. (2018b). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? *International Review of Administrative Sciences* (forthcoming).

APPENDIX

Appendix A: The 10 data quality principles of the Sunlight Foundation

	Definition
Completeness	Data sets released by the government should be as complete as possible, reflecting the entirety of what is recorded about a particular subject. All raw information from a data set should be released to the public, except to the extent necessary to comply with federal law regarding the release of personally identifiable information. Metadata that defines and explains the raw data should also be included, along with formulas and explanations for how derived data was calculated. Doing so will permit users to understand the scope of the available information and to examine each data item at the greatest possible level of detail.
Primacy	Data sets released by the government should be primary source data. This includes the original information collected by the government, details on how the data was collected, and the original source documents recording the collection of the data. Public dissemination will allow users to verify that information was properly collected and accurately recorded.
Timeliness	Data sets released by the government should be available to the public in a timely way. Whenever feasible, information collected by the government should be released as quickly as it is gathered and collected. Priority should be given to data whose utility is time-sensitive. Real-time information updates would maximise the utility the public can obtain from this information.
Ease of physical and electronic access	Data sets released by the government should as accessible as possible, with accessibility defined as the ease with which information can be obtained, whether through physical or electronic means. Barriers to physical access include requirements to visit a particular office in person or requirements to comply with particular procedures (such as completing forms or submitting FOIA requests). Barriers to automated electronic access include making data accessible only via submitted forms or systems that require browser-oriented technologies (e.g. Flash, JavaScript, cookies or Java applets). By contrast, providing an interface for users to download all of the information stored in a database at once (known as bulk access) and the means to make specific calls for data through an application programming interface (API) make data much more readily accessible. (An aspect of this is *findability*, which is the ability to easily locate and download content.)
Machine-readability	Machines can handle certain kinds of inputs much better than others. For instance, handwritten notes on paper are very difficult for machines to process. Scanning text via optical character recognition (OCR) results in many matching and formatting errors. Information shared in the widely used PDF format, for instance, is very difficult for machines to parse. Thus, information should be stored in widely used file formats that easily lend themselves to machine processing. (When other factors necessitate the use of difficult-to-parse formats, data should also be available in machine-friendly formats.) These files should be accompanied by documentation related to the format and how to use it in relation to the data.
Non-discrimination	*Non-discrimination* refers to who can access data and how they must do so. Barriers to data use can include registration or membership requirements.

	Another barrier is the uses of a *walled garden*, which is when only some applications are allowed access to data. At its broadest, non-discriminatory access to data means that any person can access the data at any time without having to identify themselves or provide any justification for doing so.
Commonly owned or open standards	Commonly owned or open standards refer to who owns the format in which data is stored. For instance, if only one company manufactures the programme that can read a file where data is stored, access to that information is dependent on use of the company's processing programme. Sometimes that programme is unavailable to the public at any cost, or is available but for a fee. For instance, Excel is a fairly commonly used spreadsheet programme that costs money to use. Freely available alternative formats often exist via which stored data can be accessed without the need for a software licence. Removing this cost makes the data available to a wider pool of potential users.
Licencing	The imposition of terms of service, attribution requirements, restrictions on dissemination and so on are barriers to public use of data. Maximal openness includes clearly labelling public information as a work of the government and available without restrictions on use as part of the public domain.
Permanence	The capability of finding information over time is referred to as permanence. Information released by the government online should be *sticky*: it should be available online in archives in perpetuity. Information is often updated, changed or removed without any indication that an alteration has been made. Or, it is made available as a stream of data, but is not archived anywhere. For best use by the public, information made available online should remain online, with appropriate version-tracking and archiving over time.
Usage cost	One of the greatest barriers to access to ostensibly publicly-available information is the cost imposed on the public for access, even when the cost is small. Governments use a number of bases for charging the public for access to their own documents: the costs of creating the information; a cost-recovery basis (cost to produce the information divided by the expected number of purchasers); the cost to retrieve information; a per page or per inquiry cost; processing cost; the cost of duplication, etc.

Appendix B: CKAN metadata fields used to create the Compliance Index

	Metadata fields
Completeness	**Original metadata** metadata_modified contact_points.e-mail resources.issued resources.download_url resources.rights organisation.name **Created metadata** Metadata existed, RawInformationExist
Primacy	**Original metadata** contact_points.e-mail
Timeliness	**Original metadata** Accrual_periodicity Modified **Created metadata** annual, semi-annual, quarterly, monthly, monthly, weekly, biweekly, daily, continuous, sum, modifiedTransformed, time until today, tau
Easy access	**Original metadata** resources.rights resources.download_url
Machine-readable format	**Original metadata** resources.format
Non-discrimination	**Original metadata** resources.rights resources.download_url resources.url
Commonly owned or open standards	**Original metadata** resources.format
Open licencing	**Original metadata** resources.rights
Permanence	**Original metadata** resources.download_url resources.url
Usage cost	**Original metadata** resources.rights

Appendix C: Question set and chaining logic

	Questions	Chaining logic
Completeness	Q1: Is the metadata complete?	If the raw information and the metadata of this resource exist = 1, else 0
Primacy	Q2: Is there an e-mail address for a contact point/support contact?	If an e-mail address to contact the originator exists = 1, else 0
Timeliness	Q3: Is the resource up to date?	If the tau of data > 0.5 = 1, else 0
Easy access	Q4: Is the data available in bulk?	If resources.download_url exists and resources.right is NonCommercialAllowed-CommercialAllowed-ReferenceNotRequired = 1, else 0
Machine-readable format	Q5: Is the resource available in machine-readable format?	If the format used is machine-readable = 1, else 0
Non-discrimination	Q6: Does people have a limited access to the resource?	If a downloadable link exists, the licence to use data is fully open and the data machine-readable = 1, else 0
Commonly owned or open standards	Q7: Is the resource in an open file format?	If the variable resources.format is filled with an open format = 1, else 0
Open licencing	Q8: Is the resource openly licenced?	If licencing information is available = 1, else 0
Permanence	Q9: Is the published resource available over time?	If a direct downloadable link exists and if it is different from the URL link = 1, else 0
Usage cost	Q10: Is the resource freely available?	If resource the resource uses an open format and open licence = 1, else 0