ELSEVIER

Comparison of metadata quality in open data portals using the Analytic Hierarchy Process

Sylvain Kubler[a,b,*], Jérérmy Robert[c], Sebastian Neumaier[d], Jürgen Umbrich[d], Yves Le Traon[c]

[a] Université de Lorraine, CRAN, UMR 7039, Campus Sciences, BP 70239, Vanduvre-lès-Nancy F-54506, France
[b] CNRS, CRAN, UMR 7039, France
[c] University of Luxembourg, Interdisciplinary Centre for Security, Reliability & Trust, 4 rue Alphonse Weicker, Luxembourg L-2721, Luxembourg
[d] Vienna University of Economics and Business, Institute for Information Business, Welthandelsplatz 1 1020, Vienna, Austria

A B S T R A C T

The quality of metadata in open data portals plays a crucial role for the success of open data. E-government, for example, have to manage accurate and complete metadata information to guarantee the reliability and foster the reputation of e-government to the public. Measuring and comparing the quality of open data is not a straight-forward process because it implies to take into consideration multiple quality dimensions whose quality may vary from one another, as well as various open data stakeholders who – *depending on their role/needs* – may have different preferences regarding the dimensions' importance. To address this Multi-Criteria Decision Making (MCDM) problem, and since data quality is hardly considered in existing e-government models, this paper develops an Open Data Portal Quality (ODPQ) framework that enables end-users to easily and in real-time assess/rank open data portals. From a theoretical standpoint, the Analytic Hierarchy Process (AHP) is used to integrate various data quality dimensions and end-user preferences. From a practical standpoint, the proposed framework is used to compare over 250 open data portals, powered by organizations across 43 different countries. The findings of our study reveals that today's organizations do not pay sufficient heed to the management of datasets, resources and associated metadata that they are currently publishing on their portal.

1. Introduction

Open data is gaining importance in the context of a growing demand for openness of public and private organizations. Organizations from all over the world are under increasing pressure to release their data to a variety of users (citizens, businesses, academics, civil servants…), leading to increased public transparency (Attard, Orlandi, Scerri, & Auer, 2015) and allowing for enhanced data-enriched public engagement in policy and other analysis (Gurstein, 2011). Data openness is expected to open up opportunities for new and disruptive digital services that potentially benefit the whole society, e.g. making specific databases easily accessible through mobile apps (Cegarra-Navarro, Garcia-Perez, & Moreno-Cegarra, 2014; Conradie & Choenni, 2014; Janssen, Charalabidis, & Zuiderwijk, 2012; Kučera, Chlapek, & Nečaskỳ, 2013).

Although opportunities are wide and worth exploring, data quality issues in open data are a crucial factor for the open data project in the long term (Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, Janssen, et al., 2012a; Kučera et al., 2013; Reiche, Höfig, & Schieferdecker,

2014). Missing metadata directly affects search and discovery services to locate relevant datasets for particular consumer needs, adding that incorrect descriptions of the datasets pose several challenges for their processing and integration with other datasets (Neumaier, Umbrich, & Polleres, 2016). The quality of the data and its description has a non-negligible impact on the reputation of the (governmental) organization publishing the data, but also on decision-making and business revenues that can be generated from open data. For example, looking at e-government benchmark frameworks, the quality of the published data is one of the key factors to be taken into consideration in the e-government assessment process (Janssen et al., 2012; Veljković, Bogdanović-Dinić, & Stoimenov, 2014), including the validation process of whether e-government goals are or not satisfied (Hernandez-Perez, Rodriguez-Mateos, Martin-Galan, & Antonia Garcia-Moreno, 2009; Jarrar, Schiuma, & Salem, 2007). High-quality data is the holy grail of any kind of policy making action as it is the sole prerequisite that can support decision making, regardless of the completeness and architectural excellence of the employed model (Ouzzani, Papotti, & Rahm, 2013). Indeed, good models perform well as long as the data they are fed with

---

Table 1
List of acronyms used throughout the article.

(RESTful) API	(REpresentational State Transfer) Application Programming Interface	AHP	Analytic Hierarchy Process
CKAN	Comprehensive Knowledge Archive Network	CSV	Comma Separated Value
CI, CR	Consistency Index, Consistency Ratio	DCAT	Data Catalog Vocabulary
IANA	Internet Assigned Numbers Authority	LOD	Linking Open Data
MCDM	Multi-Criteria Decision Making	ODPQ	Open Data Portal Quality
SME	Small and Medium-sized Enterprises	PDF	Portable Document Format
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluations	RDF	Resource Description Framework
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution	OKF	Open Knowledge Foundation
W3C	World Wide Web Consortium		

is of sufficient quality (Koussouris, Lampathaki, Kokkinakos, Askounis, & Misuraca, 2015).

Organizations and governments are well aware of the quality problems, even publishing guidelines and best-practices to improve the quality of their (meta) data. For instance, the Australian government provides a set of data quality guidelines to guarantee a certain level of quality at their portal (Waugh, 2015). At the same time, various efforts emerge to assess and monitor the quality of data portals, which supports the providers to identify and address quality issues. A good overview is presented in a white paper of the Open Data Institute (2015). In addition, we also contribute to this development with our Open Data Portal Watch framework, which makes it possible the monitoring and assessment of the quality of over 250 open data portals (Neumaier et al., 2016). Consequently, the data of such quality assessment initiatives can be used to compare portals with each other and report/justify on the effectiveness of certain quality improvement efforts. However, one of the challenges to properly compare/rank data portals lies in the task of processing multiple quality indicators, all of which may address different aspects of open data in e-government, adding that open data stakeholders may have completely different needs/preferences regarding the indicators' importance. Given the MCDM nature of the problem and evidences that there is a lack of frameworks and tools to dynamically assess the data quality in place (Veljković et al., 2014; Zuiderwijk & Janssen, 2014b), this paper presents an ODPQ web dashboard[1] that acts as a decision support tool for open data stakeholders to assess, and most importantly compare, a set of open data portals. Governmental organizations, for example, can benefit from the ODPQ dashboard to rate each other based on a common set of open data quality indicators which may, in turn, help them to perform part of the quality and quantity assessment process in e-government benchmarking exercises (Veljković et al., 2014), as will be discussed in this paper. In the same vein, the dashboard can foster collabration collaboration between organizations (e.g., to identify one or more organizations that are good, or experienced, in managing quality of open data), but also as a means to stimulate sustained efforts towards the continuous improvement of data quality (Zuiderwijk & Janssen, 2014a).

The summary of the paper is as follows: Section 2 discusses how open data stands in relation to e-government and existing quality indicators. Section 3 provides insight into the research methodology underlying the ODPQ framework development. Section 4 shows how the ODPQ dashboard can be used by open data stakeholders to monitor, assess and rank active open data portals (over 250 in this showcase showcase) according to personal needs and preferences. Conclusions, implications, limitations and future research are discussed in Section 5. All acronyms used in this article are summarized in Table 1.

2. Open data and e-government

In recent years, a number of open data movements sprung up

around the world, with transparency and data reuse as two of the major aims (Attard et al., 2015). To mention a few, there is the Public Sector Information Directive in 2003 in Europe, U.S. President's Obama open data initiative in 2009, and the G8 Open Data Charter in 2013. Open government data portals resulting from such movements provide means for citizens and stakeholders to obtain government information about the locality or country in question. In this context, open data is an integral part of open and e-government (Kučera et al., 2013), as will be discussed in Section 2.1. Section 2.2 provides a more representative picture of an open e-government model, along with literature-based evidences that open data is one of the most, if not the most, important pillars of such models. In view of our research focus, Section 2.3 discusses criteria for metadata quality assessment of open data portals in relation to the existing literature.

2.1. Relationship between open, government & linked data

Open data has truly defined an open government concept where governmental data of public interest is available without any restriction, being easily found and accessed, thus contributing to enhance public trust and confidence in governments (Tolbert & Mossberger, 2006). As discussed in Attard et al. (2015), open government data is a subset of open data and is simply government-related data that is made open to the public using an appropriated data license. Government data might contain multiple datasets, including budget and spending, population, census, geographical, parliament minutes, and so on. It also includes data that is indirectly 'owned' by public administration such as data related to climate/pollution, public transportation, congestion/traffic (Veljković et al., 2014). Several countries have already demonstrated their commitment to opening government data by joining the Open Government Partnership (Open Knowledge International, 2017). Some open data is also "linked data", which relies on the idea that the mechanisms used nowadays to share and interlink documents on the Web can be applied to share and interlink data and metadata about these documents, as well as concepts and entities they relate to Bizer, Heath, and Berners-Lee (2009). The most visible example of adoption and application of the linked data principles is the Linking Open Data (LOD) initative initiative (Attard et al., 2015).

The ODPQ framework proposed in this paper falls within the scope of (linked) open government data, whose main pillars and concepts are more thoroughly discussed in the next section based on a referenced e-government benchmark model.

2.2. Open e-government benchmark model

Various e-government benchmarks have been developed and confirmed in practice over the past decade, spanning from e-government 1.0 and 2.0 models (Baum & Di Maio, 2000; Eggers, 2007) to open government models (Lee & Kwak, 2012; Parycek & Sachs, 2010). Nonetheless, in a recent paper, Veljković et al. (2014) argued that there was no suitable open government benchmark and, accordingly, proposed a five-indicator model:

---

[1] http://mcdm.jeremy-robert.fr, accessed on Nov., 2017.