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

Sylvain Kubler, Jérémy Robert, Sebastian Neumaier, Jürgen Umbrich, Yves Le Traon

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

Sylvain Kubler[a,b,*], Jérémy 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 L-2721 Luxembourg
[d] *Vienna University of Economics and Business, Institute for Information Business*
Welthandelsplatz 1 1020 Vienna, Austria

Abstract

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 straightforward 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.

Keywords: Open Data, e-government, Data Quality, Analytic Hierarchy Process, Multi-Criteria Decision Making, Decision Support System

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 et al., 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 (Janssen et al., 2012; Kučera et al., 2013; Conradie and Choenni, 2015; Cegarra-Navarro et al., 2014).

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 et al., 2012a; Kučera et al., 2013; Reiche et al., 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 et al., 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 (Veljković et al., 2014; Janssen et al., 2012), including the validation process of whether e-government

*Corresponding author
Email addresses: s.kubler@univ-lorraine.fr (Sylvain Kubler), jeremy.robert@uni.lu (Jérémy Robert), sebastian.neumaier@wu.ac.at (Sebastian Neumaier), juergen.umbrich@wu.ac.at (Jürgen Umbrich), yves.letraon@uni.lu (Yves Le Traon)

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		

goals are or not satisfied (Jarrar et al., 2007; Hernandez-Perez et al., 2009). 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 et al., 2013). Indeed, good models perform well as long as the data they are fed with is of sufficient quality (Koussouris et al., 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 (Open Data Institute, 2016). 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 et al., 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 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 et al., 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 showhcase) 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.

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

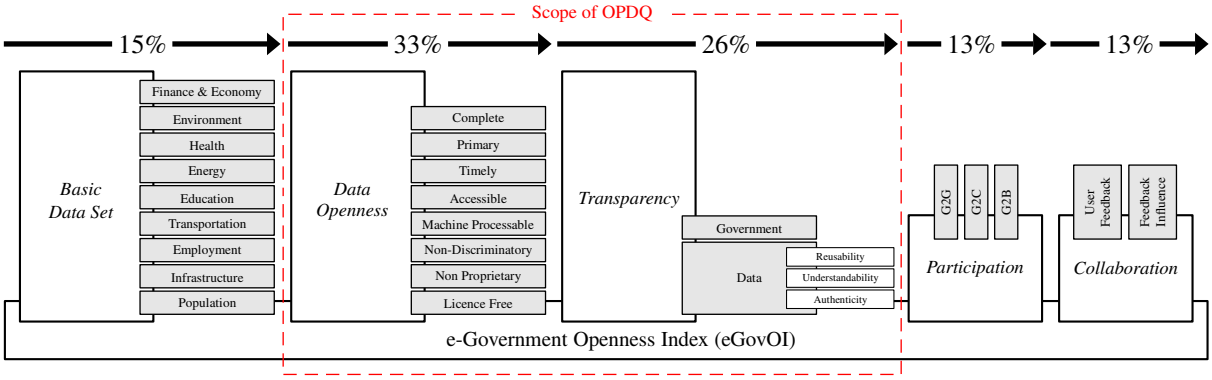


Figure 1: Details on the e-Government Openness Index (eGovOI) model proposed by Veljković et al. (2014).

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 and 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 et al., 2009). The most visible example of adoption and application of the linked data principles is the Linking Open Data (LOD) initative (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 and Di Maio, 2000; Eggers, 2007) to open government models (Parycek and Sachs, 2010; Lee and Kwak, 2012). 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. *Basic data set indicator:* determines the presence of a predefined set of high-value open data based on nine categories: Finance & Economy, Environment, Health, Energy, Education, Transportation, Employment, Infrastructure, Population;

2. *Data openness indicator:* focuses on evaluating the degree of openness of the published data based upon eight criteria that are consistent with the Open Government WG (2007)'s list of preferable characteristics for open data;

3. *Transparency indicator:* consists of two indicators (i) Government Transparency, which is observed as a measure of insight into government tasks, processes and operations; and (ii) Data Transparency, which is calculated as an average of the Authenticity, Understandability and Data Reusability values;

4. & 5. *Participation & Collaboration indicators:* user involvement is used as a source for participation and collaboration indicators.

The authors use these five indicators and underlying criteria to compute an overall index, referred to as eGovOI (e-Government Openness Index, *cf.* Figure 1), which makes it possible to monitor the progress of governments over time. Figure 1 also emphazises to what extent each of the five indicators contributes to the overall eGovOI index (e.g., Data Openness indicator has an importance of 33% with respect to the other indicators). Our research work

Table 2: Quality dimensions derived from DCAT and used in the ODPQ assessment & comparison process, see (Neumaier et al., 2016)

Dimensions	Sub-dimensions		Description	Metric
Existence (Q_e)	Access	$Q_{e(acc)}$	The extent to which access information for resources is provided	%
	Discovery	$Q_{e(dis)}$	The extent to which information helping to discover/search datasets is provided	%
	Contact	$Q_{e(con)}$	The extent to which information helping to contact the dataset owner is provided	%
	Rights	$Q_{e(rig)}$	The extent to which information about the dataset's or resource's license is provided	%
	Preservation	$Q_{e(pre)}$	The extent to which information about the resource's format, size or update frequency is provided	%
	Date	$Q_{e(dat)}$	The extent to which information about the creation and modification dates of metadata and resources is provided	%
	Temporal	$Q_{e(tem)}$	The extent to which temporal information is provided	%
	Spatial	$Q_{e(spa)}$	The extent to which spatial information is provided	%
Conformance (Q_c)	AccessURL	$Q_{c(acc)}$	The extent to which the values of access properties (HTTP, URLs) are valid	%
	ContactEmail	$Q_{c(ema)}$	The extent to which the email contact properties are valid	%
	ContactURL	$Q_{c(ext)}$	The extent to which the URL/HTTP contact properties are valid	%
	DateFormat	$Q_{c(dat)}$	The extent to which the date information is specified using a valid date format	%
	License	$Q_{c(lic)}$	The extent to which the license maps to the list of licenses given at (Open Knowledge International, 2017)	%
	FileFormat	$Q_{c(fil)}$	The extent to which the file format or media type is registered by (IANA, 1988)	%
Retrievability (Q_r)	Dataset	$Q_{r(dat)}$	The extent to which the described dataset can be retrieved by an agent	%
	Resource	$Q_{r(res)}$	The extent to which the described resource can be retrieved by an agent	%
Accuracy (Q_a)	FormatAccr	$Q_{a(for)}$	The extent to which the specified file format is accurate	%
	SizeAccr	$Q_{a(siz)}$	The extent to which the specified file size is accurate	%
Open Data (Q_o)	OpenFormat	$Q_{o(for)}$	The extent to which the file format relies on an open standard	%
	MachineRead	$Q_{o(mac)}$	The extent to which the file format can be considered as machine readable	%
	OpenLicense	$Q_{o(lic)}$	The extent to which the used license complies with the open definition	%

– *i.e., the proposed ODPQ framework* – focuses on assessing the quality of metadata of open data portals over time, thus covering a substantial part of e-government benchmark models such as eGovOI (59% = 33% + 26%).

The next section discusses in more detail the set of criteria underlying the second and third indicators in relation to the existing literature and to the quality metrics considered in the ODPQ framework.

2.3. Data Openness & Transparency indicators

Evaluating openness and transparency in e-government depends on multiple dimensions (Veljković et al., 2014; Janssen et al., 2012; Bertot et al., 2012; Huijboom and Van den Broek, 2011), the main ones being summarized by the eGovOI model (*cf.* Figure 1). Metadata of open data sets provides a useful basis for evaluating various aspects of such dimensions. For example, high-quality metadata is key for documenting results, so that they can be interpreted appropriately, searched based on what processes were used to generate them, and so that they can be understood and used by other investigators (Sugimoto, 2014; Gil et al., 2011). Unfortunately, in practice, assessing the quality of metadata information is not an easy and straightforward process; one of the major challenges lies in the lack of commonly agreed metadata representations (Zuiderwijk et al., 2012a).

To overcome this challenge, we proposed in previous research (Neumaier et al., 2016) to perform a mapping for metadata vocabulary schemas observed on different portal software (e.g., CKAN, Socrata, OpenDataSoft) to a generic scheme, which is intended as a homogenization of different metadata sources. The quality metrics derived from this generic scheme are listed and described in Table 2. These metrics are classified into five main categories: *(i) Existence* (i.e., existence of important metadata keys); *(ii) Conformance* (i.e., does the metadata information adhere to a certain format, if existing?); *(iii) Retrievability* (i.e., availability and retrievability of the metadata and data); *(iv) Accuracy* (i.e., does the information accurately describe the underlying resources?); and *(iii) Open Data* (i.e., is the specified format and license information suitable to classify a dataset as open?). All metrics listed in Table 2 focus only on metadata and shall enable an automated and scalable assessment. To put it another way, our research work does not yet include metrics that require to inspect the content of a dataset, and metrics that require a manual assessment are currently out of scope of the study.

In the following, we discuss in greater detail how the proposed categories and associated metrics align with the eGovOI's openness and transparency criteria. Such an alignment is discussed based on Table 3, where rows correspond to the eGovOI criteria and columns to our quality metrics. A two-level scale (+, ++) is used to highlight whether our metrics slightly or strongly contribute to cover the eGovOI criteria.

2.3.1. Complete

The completeness is calculated according to five features in eGovOI: *"the presence of a data meta description, the possibility of data downloading, whether the data are machine readable and whether the data are linked*

Table 3: Summary of (i) key criteria underlying Data Openness & Transparency in e-government benchmark models, and (ii) the extent to which the quality metrics underlying ODPQ meets these criteria

	Key criteria	Associated with (similar references considered)	Existence (Q_e)								Conformance (Q_c)						Retr. (Q_r)		Accu. (Q_a)		Open data (Q_o)		
			$Q_{e(acc)}$	$Q_{e(dis)}$	$Q_{e(con)}$	$Q_{e(rig)}$	$Q_{e(pre)}$	$Q_{e(dat)}$	$Q_{e(tem)}$	$Q_{e(spa)}$	$Q_{c(acc)}$	$Q_{c(ema)}$	$Q_{c(ext)}$	$Q_{c(dat)}$	$Q_{c(lic)}$	$Q_{c(fil)}$	$Q_{r(dat)}$	$Q_{r(res)}$	$Q_{a(for)}$	$Q_{a(siz)}$	$Q_{o(for)}$	$Q_{o(mac)}$	$Q_{o(lic)}$
Data Openness	Complete	*"all public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations."* (Open Government WG, 2007) *"all the information required to have the ideal data representation"* (Veljković et al., 2014)	++	++	++	++	++	++	++	++								++	+	+		++	
	Primary	*"data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms."* (Open Government WG, 2007) *"with the finest possible level of granularity, not in aggregate forms"* (Lourenço, 2015)						++						++		+			+		++	++	++
	Timely	*"data is made available as quickly as necessary to preserve the value of the data"* (Open Government WG, 2007) *"transparency in real time"* (Heald, 2012) *"timely and accurate decisions requires reliable and relevant information"* (Rojas et al., 2014)					++	++	++											+			
	Accessible	*"data is available to the widest range of users for the widest range of purposes."* (Open Government WG, 2007) *"discoverability of open data is bound to the quality of the metadata describing the data itself"* (Attard et al., 2015) *"easiness [access, navigation]"* (Lourenço, 2015)		++		+					++						++	++	+		+	+	+
	Machine processable	*"data is reasonably structured to allow automated processing."* (Open Government WG, 2007) *"three star openness level requires the use of non-proprietary format"* (Martin et al., 2013)					+									+			+	+	++	++	
	Non-discriminatory	*"the re-use of public sector documents have to be non-discriminatory for comparable categories of re-use (e.g., for commercial and non-commercial re-use)"* (Janssen, 2011) *"data is available for all to use, without requiring any registration"* (Attard et al., 2015)				++									++								++
	Non-proprietary	*"data is available in a format over which no entity has exclusive control"* (Open Government WG, 2007) *"non-proprietary is a characteristic that open data needs to have (e.g. CSV instead of Microsoft Excel)"* (Dong et al., 2016)					+									+			+	+	++	++	
	License free	*"data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable restrictions may be allowed."* (Open Government WG, 2007) *"unclear license conditions and high up-front fees may form a barrier for potential users"* (Welle Donker and van Loenen, 2017)				++									++								++
Transparency	Reusability	*"5 Star Open data scale is widely used to evaluate data reusability"* (Berners-Lee, 2010) *"government should focus less on the portal development and more on open data reusability"* (Sieber and Johnson, 2015)					+									+			+	+	++	++	+
	Understandability	*"existence of textual description, searchable tags and links for a dataset"* (Veljković et al., 2014) *"data must be easily comprehended"* (Ren and Glissmann, 2012) *"first step to improve data understandability is to provide metadata"* (Vetrò et al., 2016)		++																			
	Authenticity	*"use of a URIs aids to improve metadata and ensure authenticity"* (Attard et al., 2015) *"government should publish information about data sources on portal, and provides possibility of reviewing datasets published by a specific data source"* (Veljković et al., 2014) *"should guard the principles of authenticity and non-repudiation of data"* (Zissis and Lekkas, 2011)			++							++	++										

(meaning that a data link is available), to ease data accessibility (e.g., embed data in a custom web application, link to other data)". In this regard, all quality metrics that fall under the existence category (Q_e) can be used to assess whether all metadata descriptions are available. $Q_{o(\text{mac})}$ (openness) can also help to assess whether the format is considered as machine readable, along with the accuracy dimension that checks whether the specified file format and size are correct. However, assessing whether "links to other data" exist is currently not supported, which would require to parse the content for links.

2.3.2. Primary

The primary criterion is partially covered by the open data-related metrics (Q_o), i.e. if the file format is conform with an open or machine readable format ($Q_{c(\text{fil})}$). If so, we can consider that the data is published in a raw format. Nonetheless, we cannot assess whether the data is published in the original format or whether a transformation or aggregation operations have been performed prior to the publishing. Indeed, this would require to have knowledge about the publishing process of the data provider.

2.3.3. Timely

This criterion is partially covered by $Q_{e(\text{pre})}$ and $Q_{e(\text{dat})}$, the former checking whether there exists any update frequency information within the metadata, the latter checking whether any creation or modification date about the metadata and underlying datasets is provided. $Q_{e(\text{tem})}$ assesses whether there is any information about the time dimension of the data itself, which can also be used as an indicator about the data freshness (i.e., is it a current or historical data?). To achieve a very accurate assessment of dataset timeliness, a resource consuming data monitoring and content inspection process would need to be set up, as discussed in (Neumaier and Umbrich, 2016).

2.3.4. Accessible

$Q_{c(\text{acc})}$ reports whether the dataset can be directly downloaded by a client without any authentication. However, this metric does not cover scenarios in which a data consumer would need to manually invoke a download link.

2.3.5. Machine processable & non-proprietary

The machine readable metric ($Q_{o(\text{mac})}$) and open format one ($Q_{o(\text{for})}$) assess whether the provided data formats can be considered as non-proprietary and machine processable (e.g., using JSON or CSV rather than an unstructured text file), along with $Q_{c(\text{fil})}$ that checks whether the file format or media type is registered by the IANA (1988).

2.3.6. Non discriminatory & License free

Providing third parties with data in a usable form, without any restriction and for free, is assessed through $Q_{o(\text{lic})}$ that checks whether the provided data license is considered to be an open license according to the opendefintion.org. To cope with specific licensing situations (e.g., a license specific to a country policy), $Q_{e(\text{rig})}$ complements $Q_{o(\text{lic})}$ by identifying whether any licensing information has been provided within the metadata.

2.3.7. Reusability

The reusability criterion is partially covered by our metrics. However, we do not inspect the content of the published data, thus making it impossible to assess whether a dataset has been published following the 5 star Linked Data principles (Bizer et al., 2009). This would indeed require to inspect the content for links and verify that these links point to existing data, which would result in thousands of HTTP lookups. Nevertheless, by assessing the machine readability of the published data formats ($Q_{o(\text{mac})}$), we do already cover the first 3 principles of the 5 star model. Furthermore, an in-depth look at existing open data portals shows that only a small portion of the total amount of datasets – *only 10K datasets over a total of 10TB (from over 259 portals)* – are currently published as RDF (the 4^{th} star), most of them being published as CSV and JSON[2].

[2]JSON is also the exchange formats for many web applications and software libraries, and some guidelines (e.g., from European Data Portal even recommend to use CSV as publishing format for open data rather than JSON or RDF.

2.3.8. Understandability

The understandability criteria is hard to assess in an automated manner and, as such, is not covered by our metrics. Nevertheless, since the discovery metric ($Q_{e(\text{dis})}$) assesses the existence of keywords, titles and descriptions within the metadata, it can serve as an indication whether the content of a dataset is or not described, thus making it easier to understand. However, only a manual assessment can clearly determine for whom and to what extent the description of a dataset is understandable. For example, a dataset published and described by an expert might be easy to understand by another expert, but not by a non-expert.

2.3.9. Authenticity

The existence metric of contact information $Q_{e(\text{ema})}$, along with the conformance of the provided contact URL and email addresses ($Q_{c(\text{ext})}$, $Q_{c(\text{ema})}$) partly cover how authentic the data publisher is, and whether there is any means to contact the publisher (e.g., for feedback or question purposes). Another option would be to check whether the portal provides a direct feedback mechanism (e.g., in the form of comment fields), but unfortunately most of today's portal software frameworks do not provide such information in their API.

3. Research methodology underlying ODPQ

The research methodology underlying the ODPQ dashboard is described in this section: section 3.1 discusses the mapping process to transform platform-specific metadata information onto a generic scheme (based on which the quality metrics listed in Table 2 were derived; section 3.2 details the approach used to aggregate such metrics as well as end-user preferences in order to obtain the final ranking of the monitored open data portals.

3.1. Open data concepts & practices

Most of the current "open" data form part of a dataset that is published in open data portals, which are basically catalogues similar to digital libraries. In such catalogues, a dataset aggregates a group of data files (referred to as resources or distributions) that are available for access or download in one or more formats (e.g., CSV, PDF, Excel). To accelerate the usage of open data by citizens and developers, it is necessary to adopt an effective open data program including API interfaces with online mapping and visualization, among other features. There exist three prominent software for publishing open data: *(i)* the open source framework CKAN[3] developed by OKF; *(ii)* the commercial Socrata open data portal[4]; and *(iii)* the recent data publishing platform OpenDataSoft[5]. These software provide ecosystems to describe, publish and consume datasets (i.e., metadata descriptions along with pointers to data resources). Such portal frameworks typically consist of a content management system, some query and search features, as well as RESTful APIs to allow agents to interact with the platform and automatically retrieve metadata and data from portals.

To overcome the lack of generic, automated and scalable frameworks for assessing the quality of open data portals over time, we proposed in previous research work a mapping from vocabulary schemas observed on data portals using the three above-mentioned software onto a generic model, intended as a homogenization of different metadata sources. This mapping relies on the W3C's DCAT metadata standard (W3C, 2016), which is an RDF vocabulary including four main classes, namely `dcat:Catalog`, `dcat:CatalogRecord`, `dcat:Dataset`, and `dcat:Distribution`. Figure 2 (*cf.*, Stage 1) provides an overview of what the W3C's DCAT metadata model looks like when mapping two distinct portals with this model. The reader can also refer to (Neumaier et al., 2016) to obtain further details about the DCAT model and associated mapping. Based on the available metadata keys in the DCAT specification, the five open data quality dimensions and underlying metrics have been proposed and introduced in previous research (Neumaier et al., 2016), as summarized in Table 2, helping to measure the quality of open data portals in a generic and scalable manner. However, the aggregation of the various quality metrics, taking into consideration both the category to which they belong to and possible end-user preferences regarding those categories/metrics, leads to a MCDM problem, as will discussed in the next section.

3.2. AHP-based comparison framework

A simplistic view of the portal quality assessment and comparison process is depicted in Figure 2, which starts by crawling, collecting and mapping datasets from distinct active open data portals to the DCAT metadata standard (*cf.*, Stage 1). Stage 2 assesses each dataset based on the quality metrics listed in Table 2, which are expressed

[3] http://ckan.org, accessed on Nov., 2017.
[4] https://www.socrata.com, accessed on Nov., 2017.
[5] https://www.opendatasoft.fr, accessed on Nov., 2017.

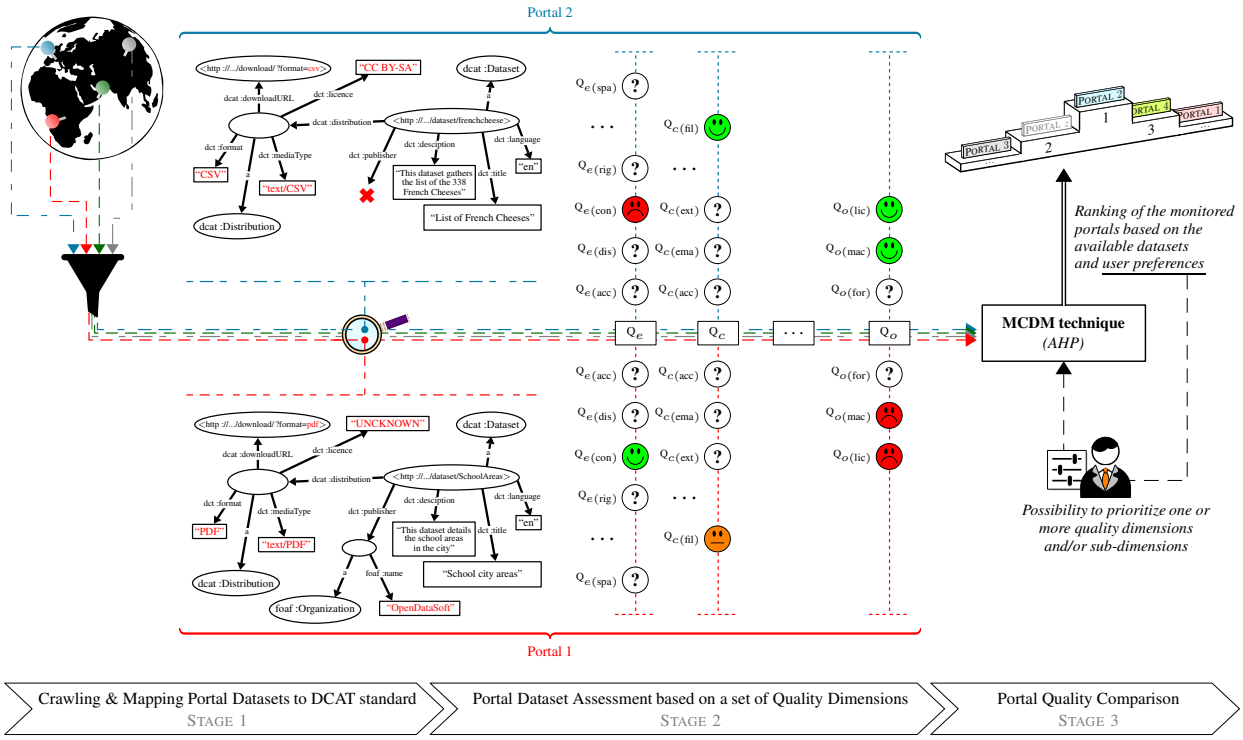


Figure 2: Overall quality assessment process: from the metadata collection to the ranking of the open data portals using a MCDM technique

as a percentage value (the higher the metric score, the higher the metadata quality). Finally, Stage 3 aggregates all the quality results and associated end-user preferences (e.g., prioritization of one or more quality dimensions) in order to obtain the final ranking of the monitored portals. So far, our research work dealt with Stages 1 and 2 (Neumaier et al., 2016). As an illustrative example, two portal datasets are considered (see Portals 1 and 2 in Figure 2). Portal 1 obtains a "good" evaluation score with respect to $Q_{e(\text{con})}$ (*cf.*, ☺ in Figure 2) since the `dct:publisher` property holds some contact information (i.e., "OpenDataSoft"), while Portal 2 does not (see ✖ and ☹). Portal 2 is nonetheless assessed positively with respect to $Q_{o(\text{mac})}$ and $Q_{c(\text{fil})}$ because (i) "CSV" is considered as a machine readable format, and (ii) both `dct:mediaType` ("text/CSV") and `dct:format` ("CSV") are registered by the IANA. Regarding Portal 1, "PDF" is not a machine readable format ($Q_{o(\text{mac})}$ evaluates to 0 for the respective dataset), however the dataset is evaluated to 0.5 with respect to $Q_{c(\text{fil})}$ (see "Neutral" smiley) because "PDF" is not a valid media type (`dct:mediaType`) but a valid format description (`dct:format`). The dataset of Portal 2 is assessed positively with respect to $Q_{o(\text{lic})}$ since `CC-BY-SA` is considered as open according to opendefinition.org, while Portal 1 is assessed negatively due to the lack of licensing information. Although not detailed here, similar examples could be elaborated regarding all the other quality metrics for which a question mark appears in Figure 2.

The MCDM nature of the problem (i.e., Stage 3), and particularly the possibility for end-users to specify their preferences about the metric priorities to obtain the final ranking of portals has not been addressed yet. There are various types of MCDM techniques in the litereture such as AHP, TOPSIS, PROMETHEE or still Fuzzy MCDM, some of them having been applied to handle e-government problems (Kubler et al., 2016a; Mardani et al., 2015). In this study, we decided to apply the AHP technique for a twofold reason: *i)* our problem deals only with linear preferences, and *ii)* AHP is an efficient and well-established technique to integrate expert knowledge, as well as tangible system properties. It should be added that AHP is, according to a recent survey (Mardani et al., 2015), the second most used MCDM technique with a frequency of application of 15.82%. AHP, originally introduced by Saaty (1977, 1980), has the advantage of organizing critical aspects of the problem in a manner similar to that used by the human brain in structuring the knowledge (i.e., in a hierarchical structure of different levels including the overall goal, the set of criteria, sub-criteria, and alternatives). The MCDM ranking problem of our study is broken down into a hierarchical structure consisting of four distinct levels:

- *Goal level:* to assess and rank the monitored open data portals in terms of published metadata quality;
- *Criteria & Sub-criteria levels:* respectively correspond to the quality dimensions and sub-dimensions given

Table 4: Tabular overview of the use case data (i.e., crawled portals and associated datasets/resources)

		CKAN	OpenDataSoft	Socrata
Number of monitored Portals		148	11	100
Number of Portals per Continent	East Asia & Pacific	10	0	0
	Europe & Central Asia	86	9	8
	Latin America & Caribbean	12	0	0
	North America	27	2	90
	South Asia	9	0	1
	Sub-Saharan Africa	4	0	1
Number of Datasets	*min*	0	0	0
	avg	4781	160	799
	max	194851	1905	10686
Number of Resources	*min*	0	0	0
	avg	15801	743	884
	max	498390	7304	28404
Number of unreachable portals per week		19.1	2.4	7.47

in Table 2. It should be noted that the hierarchical model is not perfectly balanced in our study (e.g., 7 Q_e sub-criteria *vs.* 2 Q_a sub-criteria), when one knows that unbalanced models may sometimes lead to biased results. However, we stick with this choice to fully match with the set of metrics derived from the DCAT mapping. The impact of a non-perfectly balanced model should nonetheless be evaluated and tackled in future work (e.g., re-designing the hierarchical structure or using structural adjustment techniques);

- *Alternative level:* the alternatives correspond to the set of monitored portals.

Given the AHP structure, several computational steps are performed to obtain the final ranking of alternatives with respect to the overall goal. Nonetheless, in view of the journal's scope and audience, we decided not to detail such computational steps in this paper, but the reader can refer to (Kubler et al., 2016b) to obtain more details. Indeed, even though the referenced paper focuses only on metrics specific to the CKAN software, the computational steps related to AHP remain unchanged. In the end, after applying AHP, each portal is ranked amongst the set of portals/alternatives in a relative way. Various rankings can be generated depending on the granularity of the analysis, e.g. one ranking with respect to each quality dimension or one unique ranking with respect to the overall goal, as will be detailed through the showcase presented in the following section.

4. ODPQ dashboard implementation & Results

This section presents how the ODPQ framework and associated web dashboard can be used by open data portal stakeholders (including governments, municipalities, or entrepreneurs) when performing quality and quantity assessment in e-government benchmarking exercises, or when developing innovative open-data based applications.

Figure 3 presents the overall architecture, including the "Backend systems", "Web/User Interfaces", as well as the set of interactions between the different system components (databases, portals, end-users...). The architecture differentiates the "Open Data Portal Watch" components developed in our previous work (Neumaier et al., 2016) (allowing for the collection, storage, DCAT mapping, and assessment of the portal metadata quality, *cf.* ① to ④ in Figure 3) and the ODPQ dashboard when an end-user requests for the open data portal quality comparison service (*cf.* ⑤ to ⑨). A RESTful API[6], denoted by API1 in Figure 3, makes it possible to retrieve various types of information about the monitored portals (e.g., stats including quality scores of one or more portals over a period of time). From a chronological standpoint, the ODPQ backend system retrieves – *through API1* – the computed data quality metrics in order to start the AHP-based comparison process (see ⑦). Since such comparisons are carried out at different intervals of time (e.g., on a weekly or monthly basis), we also compute the ranking and quality evolution of the portals over time (see ⑧). Similarly to API1, a second RESTful API (denoted by API2 in Figure 3) enables end-users to retrieve ranking results over specific periods of time and depending on their preferences.

The following sections focus on stages ⑤ to ⑨, having 259 open data portals monitored over 47 weeks (from week 27 2016 to week 20 2017). Table 4 summarizes the showcase data, namely (i) the distribution of the CKAN, Socrata and Opendatasoft software frameworks on the basis of the 259 monitored portals; (ii) the distribution of software per continent; (iii) the minimal, average, and maximal number of datasets and resources (per software) held by the 259 portals; as well as (iv) the average number of portals (per software framework) that were unreachable per week. One interesting finding is that CKAN is predominantly used in Europe & Central Asia (86 open

[6] http://data.wu.ac.at/portalwatch/api, accessed on Nov., 2017.

data portals), while Socrata is mostly used in the North America (90 portals). Another finding of our study is that only 12.7% of the 259 monitored portals were (in average) unreachable during the weekly crawling process, which makes us confident about the relevance of our results/findings. It should nonetheless be noted that, for practical reasons, we decided not to take into account yet the Accuracy (Q_a) and Retrievability (Q_r) dimensions in the AHP analysis because: (i) accuracy metrics require to inspect the data content to verify that the specified file format and file size in the metadata is accurate. However, due to limited resources for downloading and parsing the files, we are not performing the accuracy assessment over all portals, which prevents us from performing a fair comparison between the 259 portals; (ii) retrievability metrics require to perform HTTP lookups to check whether the content can be downloaded. The main challenge here is to perform these lookups in a reasonable amount of time. Even though a straightforward solution would be to perform HTTP Head lookups, many portals such as Socrata do not support such a protocol, preventing us once more from having a fair comparison between all portals. Such issues should be tackled in future implementation of ODPQ in order to include these quality metrics in the implemented comparison process.

The summary of the section is as follows: Section 4.1 presents the comparison results for a specific week (week 1, 2017), assuming that all criteria are of equal importance. Considering the selected week, section 4.2 shows how end-user preferences can lead to radically different rankings, which may affect subsequent decision-making. Section 4.3 gives insight into the evolution – *over almost one year (47 weeks)* – of the portal rankings and resource availability. In an effort of clarity, we use portal indexes (from 1 to 259) rather than exact names, but the reader can refer to Table A.6 to identify the matching: *Index* $\leftrightarrow$ *Portal name*.

4.1. Portal ranking (Week 1, 2017): Equivalence between criteria

The ODPQ dashboard provides end-users with a set of functionalities, enabling them to:

- vizualize the AHP hierarchy considered in the study, as shown in the dashboard screenshot annotated by ❶ in Figure 4;
- vizualize the relative quality score obtained by each open data portal for a specific week, as shown with the screenshot annotated by ❷ in Figure 4;
- vizualize the ranking of one or more portals with regard to one or more quality dimensions, making it possible to more thoroughly analyze how a portal behaves regarding the selected dimensions. This view corresponds to the screenshot annotated by ❸;
- modify his/her preferences regarding the criteria importance, e.g. if the end-user wants to give – *at a specific point in time and for specific reasons* – more importance to one dimension (e.g., Openness Q_o over Conformance Q_c) or sub-dimension (e.g., to focus more on the Format openness $Q_{o(F)}$ than on the License openness $Q_{o(L)}$). This view corresponds to the screenshot annotated by ❹ in Figure 4 (sliders corresponding to the pairwise comparisons performed at the criteria level in AHP).

In the first scenario, the end-user wants to analyze the portal rankings without prioritizing any quality dimension. Figure 5 gives insight – *in the form of a histogram* – into the quality comparison results, where the x-axis refers to the 259 portal indexes and the y-axis to the relative quality score obtained after applying AHP. It can be observed that portals 67 and 107 have the highest scores when having all criteria equal in importance.

Besides this observation, we now assume that the end-user is particularly interested in portals located in Brazil since she/he is carrying out a study on the quality of open data portals managed by brasilian institutions/organizations. As a first observation, the histogram seems to highlight that portal 22 (i.e., *dados_recife_pe_gov_br*) has the best quality among the five brazilian portals. To study more throroughtly the reason behind such a ranking/finding, the end-user uses the dashboard view ❸ (*cf.*, Figure 4), where she selects the five brazilian portals and vizualizes how they behave with respect to the three quality dimensions Q_e, Q_c, Q_o. The comparison results are given in the form of a polar chart in Figure 6 (the larger the surface area, the better the portal ranking, and consequently the metadata quality). It can be observed that the five portals are ranked among the top 100 with regard to each quality dimension, except portals 20 and 24 (i.e., *dados_al_gov_br* and *dadosabertos_senado_gov_br*) that have a poor ranking respectively regarding the open data dimension for portal 20 (ranked 191^{st}) and the Conformance dimension for portal 24 (ranked 134^{th}). The point of all this is to show that the ODPQ dashboard provides advanced features/views to help end-users to navigate through the different views and better understand why a portal has a poor (or high) ranking/quality.

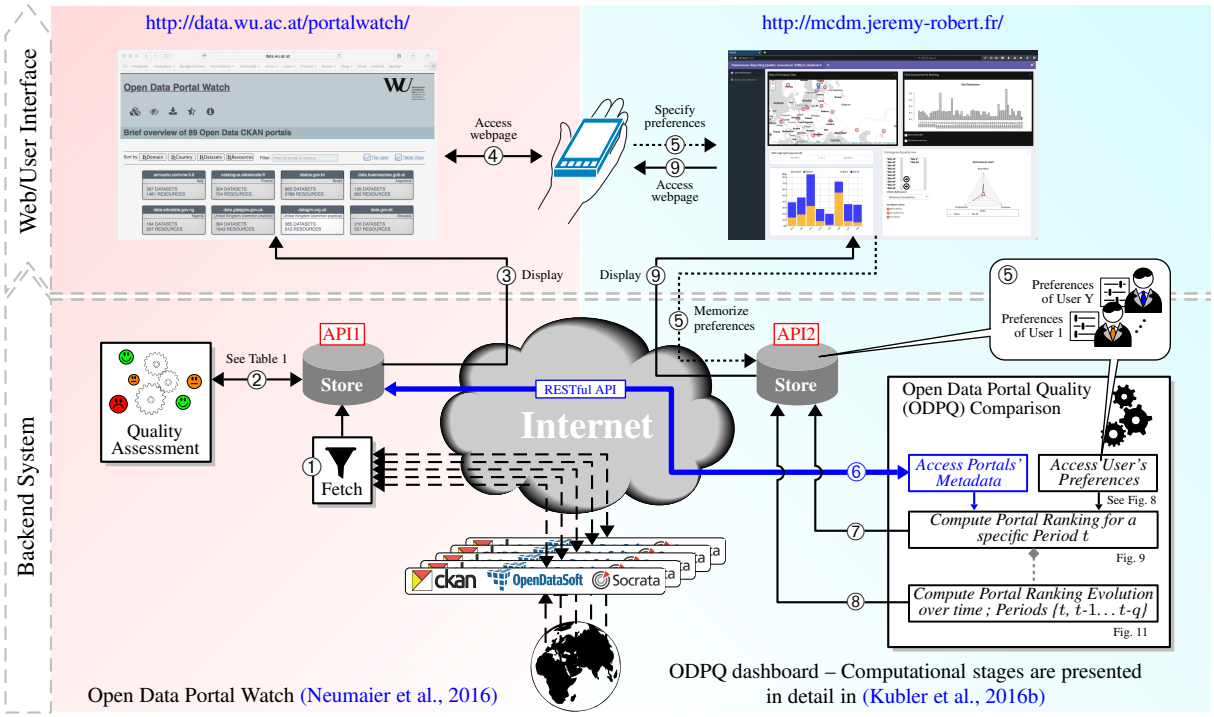


Figure 3: Overall infrastructure underlying the ODPQ web dashboard

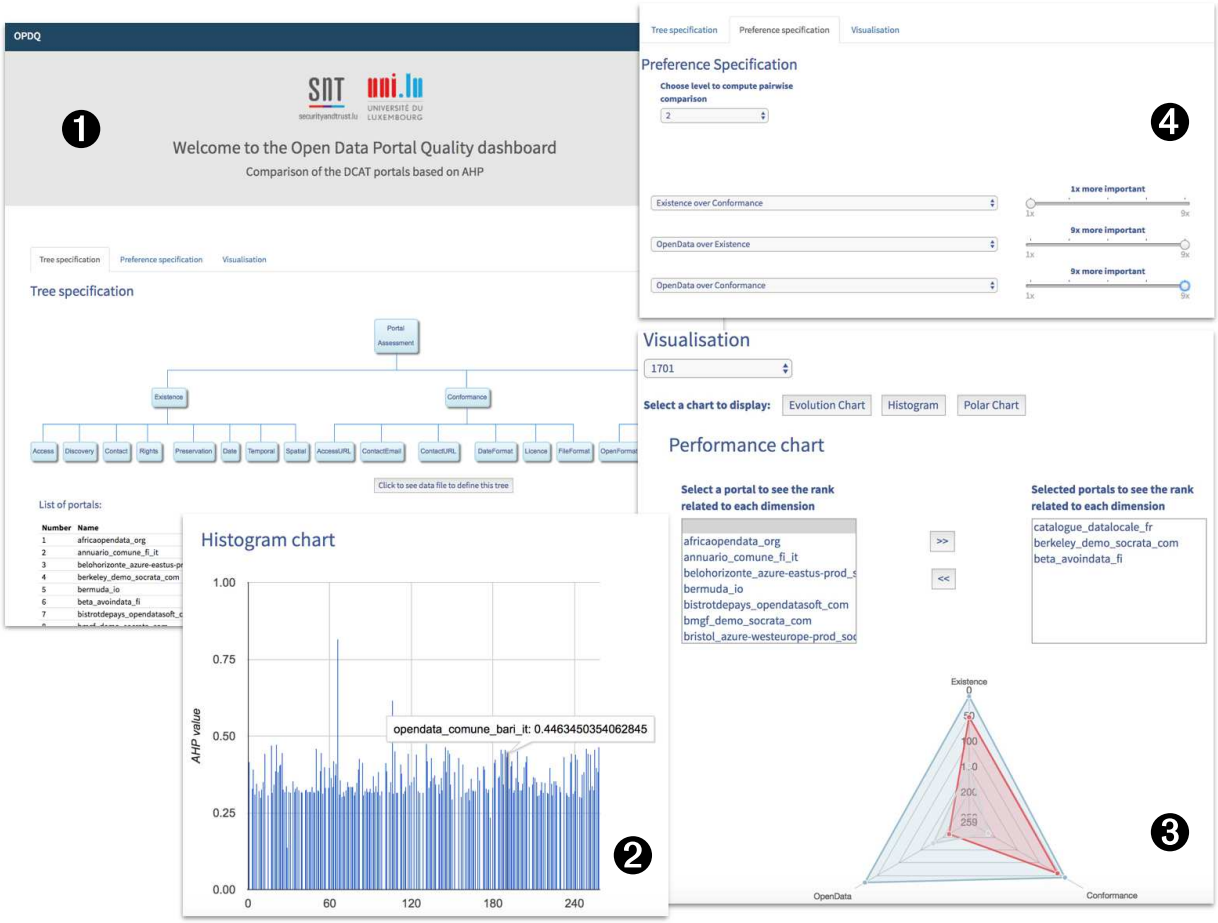


Figure 4: Screenshots of the ODPQ dashboard and associated views/functionalities

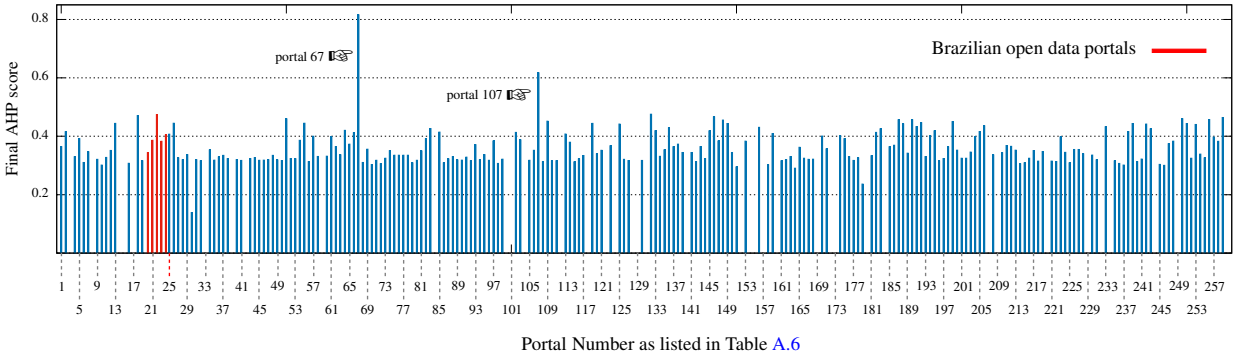


Figure 5: Final AHP scores (y-axis) obtained by the 259 open data portals (y-axis) for Week 1 (2017) – Criteria of equal importance

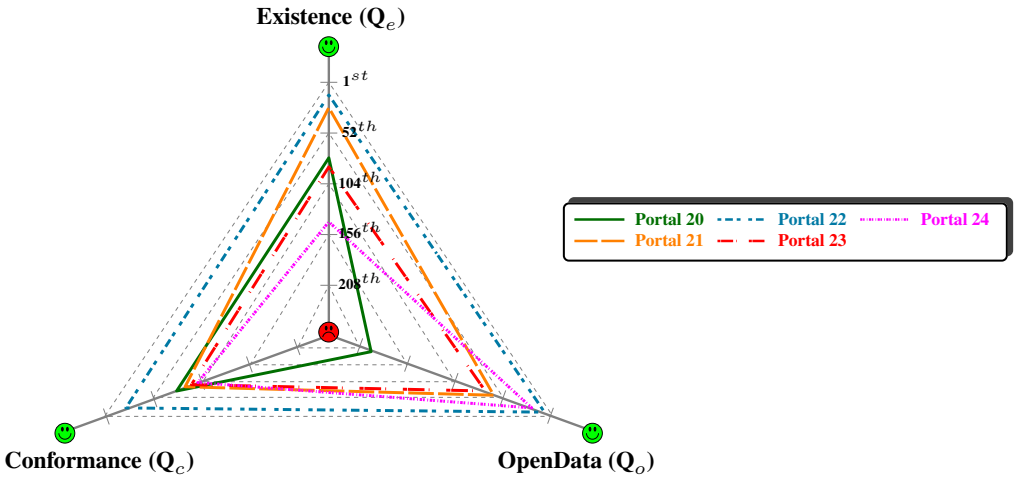


Figure 6: Brazilian open data portal comparison (Week 1, 2017)

4.2. Portal ranking (Week 1, 2017): End-user preference changes & resulting impact

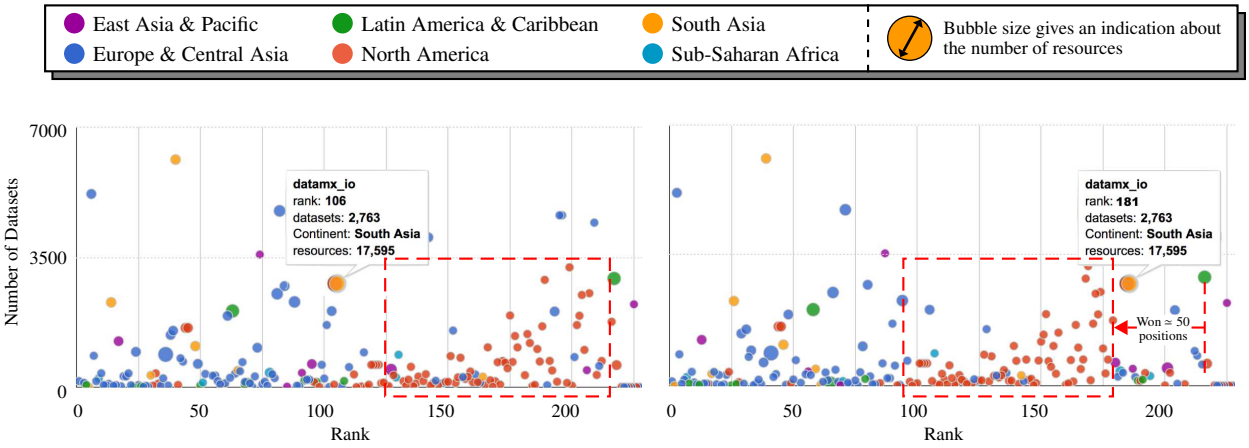
The end-user now wants to give a higher priority to the "Open Data" dimension (e.g., extreme importance over the other dimensions at level 2). To do so, the end-user uses the dashboard view ❹ presented in Figure 4.

To bring to light how the final portal ranking can be affected by end-user preferences, we propose to compare the first and second scenarios (i.e., equivalence between criteria *vs.* prioritization of open data-related metrics) taking a slightly different view in Figure 7. Each bubble refers to one specific portal (the bubble's color having been chosen according to the continent where the city portal is located/hosted), the x-axis refers to the portal indexes (from 1 to 259), the y-axis to the number of datasets held by each portal for the selected week, and the bubble size to the number of resources (the bigger the bubble, the higher the number of resources). An interesting finding is that, for equivalent preferences (see Figure 7(a)), data portals located in North America occupy the bottom of the rankings (most of them being ranked between 130-220), while the same set of portals won $\simeq$ 50 positions when prioritizing the open data dimensions (see Figure 7(b)). Even though it appears that most of the portals from the other continents remain better, this shows that the licensing on portals that have slipped down the overall rankings is less well managed than the ones located in North America. Overall, the results/rankings must be carefully studied and interpreted depending on the specified preferences.

4.3. Portal evolution over one year

The previous two sections mainly discussed the features and widgets offered by the ODPQ dashboard, and how open data stakeholders can benefit from them to make better decisions (i.e., easily adjusting the criteria importance as they see fit). However, the focus was on the comparison of open data portals for a specific week (week 53 to be precise), and not on how these portals evolve over time. This section discusses such an evolution both regarding the portal rankings (a portal can win or lose positions from week to week) and the resources held by each portal (datasets and/or resources can be deleted or added on portals).

Figure 8 provides an overview of the ranking evolution in the form of a decile boxplot (the 1^{st} and 9^{th} decile being displayed). The x-axis still refers to the portal indexes (1 to 259), while the y-axis refers to the number



(a) Equivalence between all (sub)-quality dimensions/metrics

(b) Prioritization of "OpenData" -related dimensions/metrics

Figure 7: Evolution of ranking *vs.* datasets at Week 1 (2017) having different user preferences about the importance of criteria

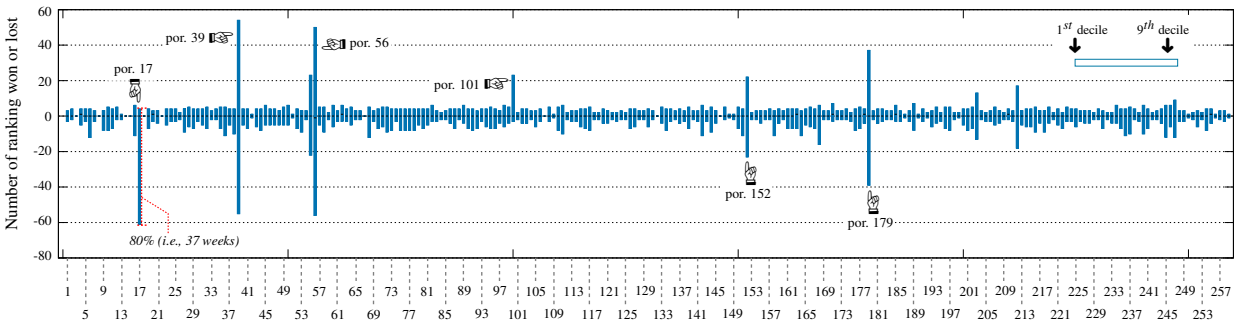


Figure 8: Overview of the deviation of portals' ranking from one week to another

of ranks that each open data portal won or lost on a weekly basis. For example, looking at portal 17, in 80% of the cases (i.e., during 37 weeks out of 47) it lost from 1 to 61 positions (see 1^{st} decile's value) and won up to 4 positions (see 3^{st} decile's value). As a result, the portal lost more than 61 positions during 5 weeks and, similarly, won more than 4 positions during 5 weeks. Although we implemented a mitigation strategy[7] to avoid a "yo-yo" effect when portals become inaccessible from one week to another (i.e., winning and loosing a high number of ranks), we observe that a few portals such as portals 39, 56, and 179 (*cf.*, Figure 8) are nonetheless affected by this effect. This is due to the fact that these portals are accessible but no datasets are available for the monitored week (may be due to maintenance operations), thus impacting on the other dimensions and leading to their downgrading in the final ranking. However, this effect is observed only for 6 portals out of the 259, which does not call into question the findings of our study. After investigation, the deviation of portals 17 and 100 is due to the addition or deletion of datasets/resources. Looking at such deviation patterns can help us to better understand the reasons of an upgrade or downgrade of a portal. Overall, and as a general comment, it can be stated that the ranking of the vast majority of portals does not evolve much (between 1 to 10 positions), which reflects to some extent the fact that governmental organizations do not pay sufficient heed in upgrading their portal's datasets.

To bring further evidence to support this statement, let us look at the resource deviation in Table 5, which provides the list of data portals that lost or won a significant number of resources from week to week (somehow reflecting the portal activity over time). Four ranges have been reported, namely portals that lost or won between [0; 10.000[, [10.000; 25.000[, [25.000; 100.000[and [100.000; 500.000[resources. Even though a few portals such as *data_gov* and *www_data_gc_ca* lost a significant number of resources ([100.000; 500.000[), we can observe that there is, in general, little activity as most of them lost/won less than 10.000 resources. To be more precise, 83% of these portals lost less than 1.000 resources, while 98% won less than 1.000 resources. This finding (i.e., little portal activity) is not a revelation for open data scholars and practitioners. Indeed, the intended positive effects and

[7]The plan consists to take the last available values related to all criteria, while downgrading the portal's accessibility dimension ($Q_{e(acc)}$).

Table 5: Overview of the deviation (from week to week) of resources held by portals

	Max. lost resources	Max. gained resources
[100.000; 500.000[	data_noaa_gov_dataset; data_gov; www_data_gc_ca; transparenz_hamburg_de	transparenz_hamburg_de
[25.000; 100.000[	data_gov_au; data_gov_uk; open-data_europa_eu; geothermaldata_org; datameti_go_jp_data_; datahub_io	data_gov; geothermaldata_org; data_gov_au
[10.000; 25.000[	opendata_socrata_com; datamx_io; datos_codeandomexico_org; edx_netl_doe_gov; data_overheid_nl; dati_trentino_it; data_hdx_rwlabs_org	edx_netl_doe_gov; data_overheid_nl; dados_rs_gov_br
[0; 10.000[	*All other portals*	*All other portals*

creating value from using open data on a large scale is easier said than done, and using open data still encounters various socio-technical impediments (Janssen et al., 2012; Zuiderwijk et al., 2012a). Although most countries legitimise their open data study based on general and macro-economic studies (e.g., Gartner, Acil Tasman...), many policy makers recognize that the precise economic impact of open data for their country remains largely unclear (Huijboom and Van den Broek, 2011). This is, from our perspective, an understandable reason why governments and other organizations do not pay sufficient heed to (i) the management of their open data portal, thus hampering the continuous feeding of portals with up-to-date datasets/resources, and (ii) the implementation of strategies to assess and compare the quality of their portal with other peer portals/organizations. ODPQ-like dashboards can be beneficial for (governmental) organizations to help them designing/building up such strategies, and stimulate them to continously improve the quality of the data they are exposing/publishing.

Before concluding this section, it is important to realise that AHP enables the comparison of alternatives, leading to a "relative" ranking of alternatives. To put it simply, it is not because a portal is ranked 1^{st} that it necessarily has a good quality; it only means that all the other alternatives/portals have a lower quality than this portal. As will be more thoroughly discussed in the conclusion section, the "absolute" measurement (Saaty, 1986) could better suit the ODPQ problem, as this approach considers a standard with which to compare elements. However, to the best of our knowledge, such a standard does not exist to date. So far, to determine whether a portal has or not a good quality, it is necessary to look at the "raw" quality metric values (expressed as a percentage in Table 2). In an effort to provide an at a glance and overall view of the "raw" quality of the 259 monitored portals, we have computed and displayed in Figure 9 the average quality score of all portals, over all weeks, with respect to each quality metric. First, it seems that the vast majority of portals obtained a very good quality score (i.e., $\geq 75\%$) regarding (i) two of the Conformance metrics, namely $Q_{c(\text{acc})}$ and $Q_{c(\text{dat})}$ respectively having valid access properties and date formats, and (ii) one of the Existence metrics, namely $Q_{e(\text{con})}$ having contact information about the dataset owner. On the opposite, the monitored portals completely failed over the year to include spatial and temporal information in the metadata (see $Q_{e(\text{tem})}$ and $Q_{e(\text{spa})}$), but also to have valid URL/HTTP contact properties (see $Q_{c(\text{ext})}$). We can also add that, even though file formats appear to comply with open and machine readable formats ($Q_{c(\text{fil})}$, $Q_{o(\text{for})}$ and $Q_{o(\text{mac})}$ having an average quality score between 50% and 75%), much more remains to be done to make licenses compliant with open license formats[8] ($Q_{o(\text{lic})}$ having an average quality score of $\simeq 25\%$).

5. Conclusion, implication and future research

5.1. Conclusion

Ever more governments around the world are defining and implementing "open data" strategies in order to increase transparency, participation and/or government efficiency. The commonly accepted premise underlying these strategies is that the publishing of government data in a reusable format can strengthen citizen engagement and yield new innovative businesses. Not only should data be published, but they should actively be sought for knowledge on how to improve the government. The publication of data could have far-reaching effects both on *e*-government implementation strategies and on the public sector. In this respect, tools for monitoring and assessing the quality in the metadata and data source of open data portals are required. This is all the more true as poor data quality can hinder business decisions and government oversight efforts.

The literature review carried out in this paper brings to light the fact that there is still research to be done in the e-government domain to enable automated and scalable assessment as well as comparison of open data portal quality. This is all the more challenging because there exist several portal software frameworks on the market, leading to a 'non-uniform' publication of open data sets. To address this lack of solution, we present

[8] Based on the Open Definition: http://licenses.opendefinition.org/licenses/groups/all.json, accessed on Nov., 2017.

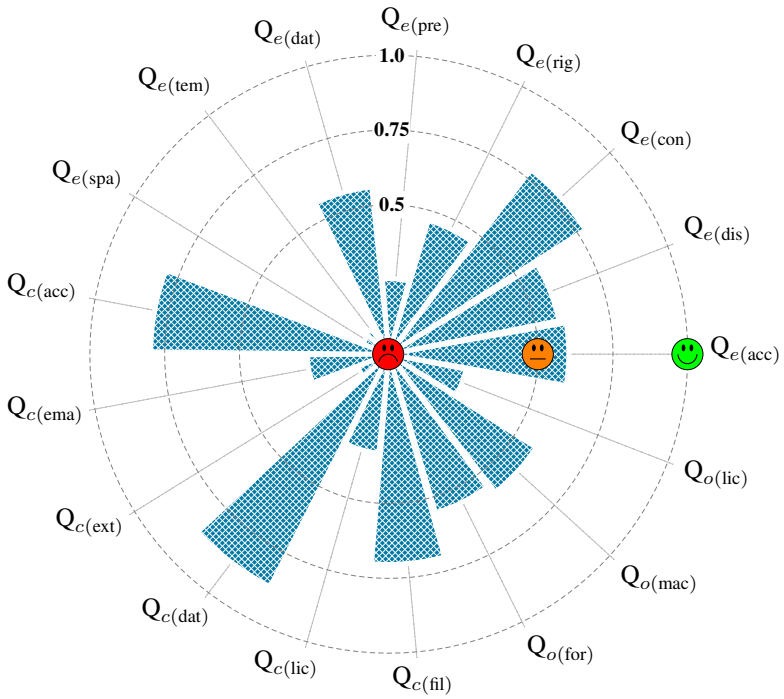


Figure 9: Average – *not "Relative"* – data quality of all open data portals with respect to each quality sub-dimension (*cf.*, Table 2)

in this paper an Open Data Portal Quality (ODPQ) dashboard, which is dynamic and enables any open data end-user/stakeholder to easily assess/rank open data portals based on multiple quality dimensions and personal preferences. Our research work purely analyzes the state and quality of the metadata, providing useful quality indicators for applications that use the metadata such as in (Tygel et al., 2016; Zuiderwijk et al., 2016, 2012b). From a theoretical standpoint, AHP is used to properly deal with such multiple indicators, while enabling end-users to adjust their preferences regarding the one or more of these indicators. This is key considering the wide range of open data stakeholders, which include:

- *upstream groups:* who supply data to the industry such as data generators and publishers (typically governments or government agencies);
- *midstream groups:* including platform developers, governments representatives involved in the role of creating an enabling environment for the practice of open data, as well as the promoters of open data;
- *downstream groups:* including data analysts, researchers, data journalists or App developers.

The proposed ODPQ framework is currently applied to assess and compare over 250 open data portals, powered by organizations across 43 different countries. A showcase is provided in this paper, which is intended to be both *(i) descriptive:* to show how easy and flexible the ODPQ dashboard can act as a decision support tool; and *(ii) analytical:* to analyze and discuss the quality of the monitored open data portals over around one year. This analysis reveals that today's organizations do not pay sufficient heed to the management of their dataset and resource descriptions. In this respect, the proposed ODPQ dashboard may prove to be of great support for organizations and policy makers to enable them to assess their portal in terms of quality, while positioning themselves with respect to peer organizations based on personal preferences. For example, a government portal typically has a strong focus on "openness" and "discoverability", while "conformance" might be of less importance. In contrast, portals hosting datasets from non-governmental organizations (e.g., the Humanitarian Data eXchange portal[9]) rather focus on the "discoverability" and "existence" dimensions. Overall, and as already discussed in this paper, the ODPQ dashboard can be of particular benefit for such organizations when performing quality and quantity assessment in benchmarking exercises, or when adopting cognitive orientation methodologies as the one recently proposed by (Moreno-Jiménez et al., 2014) for public administrations.

[9]See portal 72: https://data.humdata.org/, accessed on Nov., 2017.

5.2. Implication

The quality assessment and comparison process allows portal providers such as governmental organizations to get an overview about their data and especially to which extent their datasets are described. This directly helps to identify potential problems for the adoption and use of their data. For instance, the "existence" dimension helps to identify important missing metadata such as the license or content format. The "conformance" metrics help to identify how homogeneous the datasets are described with respect to standard formats. Overall, the primary focus of this work is on providing a metric tailored comparison of open data portals using AHP.

Nevertheless, our study also reveals some global trends for the various quality aspects of portal metadata/descriptions of datasets, as well as some limitations of our framework with regard to the data openness and transparency dimensions in e-government benchmark models. Indeed, systems such as the Open Data Barometer[10] and Open Data Portal Watch (ODPQ) can assess certain quality aspects of portals and allow to compare them, but they either use quality metrics that can be manually computed or metrics that make the assessment automatic and scalable. Both approaches have their advantages and disadvantages. The automatic approach provides frequent quality reports (e.g., on a weekly basis) but cannot easily integrate human knowledge about a specific portal. Also the inspection of the data content is very resource consuming, considering that 2 million resources are today available over the 250 monitored portals. The manual approach makes it possible to incorporate human background knowledge to in-depth analyze metrics such as "Understandability" (*cf.*, section 2.3), but unfortunately this is a time consuming process and is typically done on a yearly basis.

5.3. Recommendations

From a recommendation viewpoint, we would advise portal providers to establish their own set of tailored quality metrics (e.g., using the AHP-enabled preference specification feature). One the one hand, this would allow them to react effectively and preemptively to potential quality issues in the creation process of datasets (e.g., making metadata keys mandatory or suggesting values for empty ones), but also to put in place a monitoring system to gain immediate insights about the overall quality of their metadata. In addition, portal providers could also establish and assess metrics about the content of their data, potentially incorporating background knowledge about the publishing process. On the other hand, data consumers can use the quality metrics as filters in their search and discovery process, or react to quality changes of a dataset (e.g. if the quality falls below a specified threshold, they might want to discard the dataset).

We observe in our framework that the heterogeneity of the metadata description is one of the main challenges to provide general quality metrics. As such, we compute our metrics over the mapping of the metadata to DCAT. Doing so, we observe that many datasets do not provide standardized description fields for geospatial and temporal properties about the datasets' content. Also, many portals have free form fields to specify the format and license, often resulting in only partially machine understandable descriptions. Similarly, keywords and descriptions are again provided as free form fields, leading again to the challenge of mapping the terms to known concept hierarchies such as DBpedia, Yago or WikiData.

Overall, our recommendations for portal providers is to interfere more in the creation process of datasets at their portal, by:

- providing a schema/ontology/model for their metadata that maps to standards such as DCAT or DCAT-AP (DCAT Application Profile for data portals in Europe);
- deriving metadata values directly from the data in an automated way (e.g., file size, format, availability);
- restricting certain metadata values to a predefined list of options (e.g., for license descriptions, field formats);
- checking/validating the conformance of certain metadata values (e.g., URLs, emails).

By doing so, the portal can guarantee a certain quality level and also the compatibility with standards, which, in return, tremendously increase the reusability and discoverability of the data.

As discussed in section 2.3, there are also many papers referring to the 5 star Linked Data principles. However, we observe from the data, as well as from recommendations about data formats of portals, that open data is mainly published as 3 star data (being open machine readable formats such as CSV or JSON). The reasons for this is that there exists many tools and interfaces to publish data in such formats (e.g., Excel exports, JSON data structures) and also many data processing libraries natively supporting JSON, CSV or XML rather than RDF. Understanding

[10] http://opendatabarometer.org/, accessed on Nov., 2017.

the RDF data model and Linked Data in itself is fairly straightforward but the creation of Linked Data is quite challenging: (i) one has to firstly model the data in form of a graph, (ii) next search and ideally use existing vocabularies or create a new ontology for the data modelling, and (iii) one may eventually need to discover URIs in external datasets, but this typically requires the knowledge about third-party Linked Data datasets.

5.4. Limitations of the study & Research perspectives

The set of quality indicators considered in our study are applied to enable large scale and periodic monitoring tasks over multi-lingual data. That being said, these indicators are not yet sufficient to display a complete picture of a dataset's quality and usage (e.g., a data publisher and/or consumer might be interested to know to what extent a dataset is used by third parties). This relates to "reputation" metrics, or "Participation & Collaboration" metrics from the eGovOI model perspective (*cf.,* Figure 1). Reporting such information, however, requires logs and download statistics that are in general not accessible or considered in our framework. Another aspect that our metrics do not fully capture is whether key government datasets are or not published as open data (e.g., government expenditures or online access to national laws and statues). From the eGovOI perspective, this corresponds to the "Basic Dat Set" indicator. Although existing initiatives such as the Open Data Barometer and Open Data Index[11] are an attempt to assess – *on a yearly-basis* – to what extent open data is published and used for accountability, innovation or social impact, such efforts still rely on metrics that require manual assessment (e.g., call for reports, providing survey forms, *etc.*). This way of proceeding (i.e., manual assessment and additional background knowledge) inevitably leads to more subjective quality scores, adding that it prevents from carrying out large scale assessment analyses, as targeted by our ODPQ framework. Given this situation, we believe that there is still research to be done to solve this *dilemma*, i.e. making it possible to perform automated/large scale assessment tasks considering the whole e-government lifecycle, including "Participation & Collaboration"- and "Basic Dat Set"-like indicators.

A second research perspective is to tackle the problem of unbalanced hierarchical model (as discussed in section 3.2), but also to handle vagueness in decision maker judgments and above all uncertainties in the computed quality metrics. Indeed, most of the quality metrics can be modeled under uncertainty because they are computed over datasets for which the relevant information is available. For example, a license is considered as open, non-open or unknown according to opendefinition.org. Such an unknown situation could be modeled under a certain level of uncertainty using Fuzzy AHP-like methods (Kubler et al., 2016a). Another improvement of our approach would be to investigate the use of the "absolute" measurement methodology in AHP instead of the "relative" one (Saaty, 1986), the reason being twofold: (i) it is best suited to MCDM problems with a high number of alternatives; (ii) it implies to compare AHP elements with a "standard", which is more stable compared with the relative measurement methodology. However, to the best of our knowledge, such a standard has not been proposed yet in the literature, even though this would be a great contribution to the field.

Finally, as previously discussed, one interesting research topic can be how to develop automatic and scalable e-government benchmark frameworks that are able to integrate human background knowledge in the computation of metrics requiring manual inputs (e.g., 'Understandability" like metrics). The automatic computation of such metrics could eventually rely on – *and combine* – techniques such as natural language processing and ontology-based knowledge representations. To this end, open data published as RDF would make such research developments easier, but paradoxically is currently not the ideal way to go as most of today's open data is published following the 3 star data.

6. Acknowlegement

The research leading to this publication is supported by the EU's H2020 Programme for research, technological development and demonstration (grant 688203), as well as the Austrian Research Promotion Agency (grant 849982).

Appendix A. Matching of open data portal indexes and respective name/URL

Table A.6: Open data portal: indexes ↔ name/URL

[11] global.survey.okfn.org, accessed on Nov., 2017.

N°	Portal name/URL		
1	africaopendata_org	2	annuario_comune_fi_it
3	belohorizonte_azure-eastus-prod_socrata_com	4	berkeley_demo_socrata_com
5	bermuda_io	6	beta_avoindata_fi
7	bistrotdepays_opendatasoft_com	8	bmgf_demo_socrata_com
9	bristol_azure-westeurope-prod_socrata_com	10	bronx_lehman_cuny_edu
11	bythenumbers_sco_ca_gov	12	catalogodatos_gub_uy
13	catalogue_datalocale_fr	14	cdph_data_ca_gov
15	ckan_gsi_go_jp	16	ckan_odp_jig_jp
17	ckan_okfn_gr	18	ckanau_org
19	controllerdata_lacity_org	20	dados_al_gov_br
21	dados_gov_br	22	dados_recife_pe_gov_br
23	dados_rs_gov_br	24	dadosabertos_senado_gov_br
25	danepubliczne_gov_pl	26	dartportal_leeds_ac_uk
27	data_acgov_org	28	data_act_gov_au
29	data_albanyny_gov	30	data_atf_gov
31	data_austintexas_gov	32	data_baltimorecity_gov
33	data_bris_ac_uk_data_	34	data_buenosaires_gob_ar
35	data_burlingtonvt_gov	36	data_cdc_gov
37	data_cityofboston_gov	38	data_cityofchicago_org
39	data_cityofdeleon_org	40	data_cityofmadison_com
41	data_cityofnewyork_us	42	data_cityofsantacruz_com
43	data_cityoftacoma_org	44	data_cms_hhs_gov
45	data_colorado_gov	46	data_ct_gov
47	data_culvercity_org	48	data_datamontana_us
49	data_dcpcsb_org	50	data_edmonton_ca
51	data_edostate_gov_ng	52	data_eindhoven_nl
53	data_energystar_gov	54	data_glasgow_gov_uk
55	data_go_id	56	data_gov
57	data_gov_au	58	data_gov_bf
59	data_gov_gr	60	data_gov_hk_en_
61	data_gov_hr	62	data_gov_ie
63	data_gov_md	64	data_gov_ro
65	data_gov_sk	66	data_gov_uk
67	data_graz_gv_at	68	data_grcity_us
69	data_gv_at	70	data_hartford_gov
71	data_hawaii_gov	72	data_hdx_rwlabs_org
73	data_honolulu_gov	74	data_iledefrance_fr
75	data_illinois_gov	76	data_illinois_gov_belleville
77	data_illinois_gov_champaign	78	data_illinois_gov_rockford
79	data_kcmo_org	80	data_kingcounty_gov
81	data_kk_dk	82	data_ktn_gv_at
83	data_lexingtonky_gov	84	data_linz_gv_at
85	data_london_gov_uk	86	data_maryland_gov
87	data_medicare_gov	88	data_michigan_gov
89	data_mo_gov	90	data_montgomerycountymd_gov
91	data_murphytx_org	92	data_nfpa_org
93	data_nhm_ac_uk	94	data_nj_gov
95	data_noaa_gov_dataset	96	data_nola_gov
97	data_nsw_gov_au	98	data_ny_gov
99	data_oaklandnet_com	100	data_ohouston_org
101	data_ok_gov	102	data_opencolorado_org
103	data_openpolice_ru	104	data_openva_com
105	data_oregon_gov	106	data_ottawa_ca
107	data_overheid_nl	108	data_providenceri_gov
109	data_qld_gov_au	110	data_raleighnc_gov
111	data_redmond_gov	112	data_rio_rj_gov_br
113	data_sa_gov_au	114	data_salzburgerland_com
115	data_seattle_gov	116	data_sfgov_org
117	data_somervillema_gov	118	data_southbendin_gov
119	data_stadt-zuerich_ch	120	data_surrey_ca
121	data_tainan_gov_tw	122	data_taxpayer_net

Continued on next column

Continued from previous column

N°	Portal name/URL	N°	Portal name/URL
123	data_ug	124	data_undp_org
125	data_upf_edu_en_main	126	data_vermont_gov
127	data_wa_gov	128	data_weatherfordtx_gov
129	data_wellingtonfl_gov	130	data_winnipeg_ca
131	data_wokingham_gov_uk	132	data_wu_ac_at
133	data_zagreb_hr	134	datacatalog_cookcountyil_gov
135	dataforjapan_org	136	datagm_org_uk
137	datahub_io	138	datameti_go_jp_data_
139	datamx_io	140	datapilot_american_edu
141	dataratp_opendatasoft_com	142	daten_rlp_de
143	dati_lazio_it	144	dati_lombardia_it
145	dati_toscana_it	146	dati_trentino_it
147	dati_veneto_it	148	datos_alcobendas_org
149	datos_argentina_gob_ar	150	datos_codeandomexico_org
151	datos_gob_mx	152	datosabiertos_ec
153	datosabiertos_malaga_eu	154	datospublicos_org
155	donnees_ville_montreal_qc_ca	156	donnees_ville_sherbrooke_qc_ca
157	dot_demo_socrata_com	158	drdsi_jrc_ec_europa_eu
159	edx_netl_doe_gov	160	exploredata_gov_ro
161	finances_worldbank_org	162	gavaobert_gavaciutat_cat
163	geothermaldata_org	164	gisdata_mn_gov
165	govdata_de	166	hampton_demo_socrata_com
167	health_data_ny_gov	168	healthdata_nj_gov
169	healthmeasures_aspe_hhs_gov	170	hubofdata_ru
171	iatiregistry_org	172	inforegio_azure-westeurope-prod_socrata_com
173	irs_demo_socrata_com	174	leedsdatamill_org
175	linkeddatacatalog_dws_informatik_uni-mannheim_de	176	nats_demo_socrata_com_login
177	nycopendata_socrata_com	178	offenedaten_de
179	open-data_europa_eu	180	open_nrw
181	open_whitehouse_gov	182	opencolorado_org
183	opendata_aberdeencity_gov_uk	184	opendata_admin_ch
185	opendata_aragon_es	186	opendata_awt_be
187	opendata_ayto-caceres_es	188	opendata_bayern_de
189	opendata_brussels_be	190	opendata_caceres_es
191	opendata_cnmc_es	192	opendata_comune_bari_it
193	opendata_go_ke	194	opendata_go_tz
195	opendata_government_bg	196	opendata_hu
197	opendata_lasvegasnevada_gov	198	opendata_lisra_jp
199	opendata_opennorth_se	200	opendata_paris_fr_opendatasoft_com
201	opendata_rubi_cat	202	opendata_socrata_com
203	opendata_swiss	204	opendatacanarias_es
205	opendatadc_org	206	opendatagortynia_gr
207	opendatahub_gr	208	opendatareno_org
209	opengov_es	210	openresearchdata_ch
211	opingogn_is	212	oppnadata_se
213	parisdata_opendatasoft_com	214	performance_chattanooga_gov
215	performance_smcgov_org	216	performance_westsussex_gov_uk
217	pod_opendatasoft_com	218	portal_openbelgium_be
219	public_opendatasoft_com	220	publicdata_eu
221	rdw_azure-westeurope-prod_socrata_com	222	reportcard_santamonicayouth_net
223	rs_ckan_net	224	scisf_opendatasoft_com
225	stat_cityofgainesville_org	226	tourisme04_opendatasoft_com
227	tourisme62_opendatasoft_com	228	transparenz_hamburg_de
229	udct-data_aigid_jp	230	westsacramento_demo_socrata_com
231	wfp_demo_socrata_com_login	232	www_amsterdamopendata_nl
233	www_civicdata_io	234	www_criminalytics_org
235	www_dallasopendata_com	236	www_data_gc_ca
237	www_data_go_jp	238	www_data_vic_gov_au
239	www_datagm_org_uk	240	www_daten_rlp_de
241	www_dati_friuliveneziagiulia_it	242	www_datos_misiones_gov_ar
243	www_edinburghopendata_info	244	www_europeandataportal_eu
245	www_hri_fi	246	www_metrochicagodata_com

Continued on next column

Continued from previous column

N°	Portal name/URL	N°	Portal name/URL
247	www_nosdonnees_fr	248	www_odaa_dk
249	www_offene-daten_me	250	www_opendata-hro_de
251	www_opendata_provincia_roma_it	252	www_opendataforum_info
253	www_opendatamalta_org	254	www_opendatanyc_com
255	www_opendataphilly_org	256	www_opendataportal_at
257	www_opengov-muenchen_de	258	www_rotterdamopendata_nl
259	www_yorkopendata_org		

References

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

Baum, C. and Di Maio, A. (2000). Gartner's four phases of e-government model. *Gartner Group*, 12.

Berners-Lee, 5-Star Open Data. (2010). Retrieve from http://5stardata.info/en/ (accessed on Nov. 2017).

Bertot, J. C., McDermott, P., & Smith, T. (2012). Measurement of open government: Metrics and process. *45th Hawaii International Conference on System Science*, Hawaii (USA) (pp. 2491–2499).

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.

Cegarra-Navarro, J.-G., Garcia-Perez, A., & Moreno-Cegarra, J. L. (2014). Technology knowledge and governance: Empowering citizen engagement and participation. *Government Information Quarterly*, 31(4), 660–668.

Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*, 31, S10–S17.

Dong, H., Singh, G., Attri, A., & El Saddik, A. (2016). Open Data-Set of Seven Canadian Cities. *IEEE Access*, 5, 529–543.

Eggers, W. D. (2007). *Government 2.0: Using technology to improve education, cut red tape, reduce gridlock, and enhance democracy*. Rowman & Littlefield.

Främling, K., Kubler, S., & Buda, A. (2014). Universal Messaging Standards for the IoT from a Lifecycle Management Perspective. *IEEE Internet of Things Journal*, 1(4), 319–327.

Gil, Y., Szekely, P., Villamizar, S., Harmon, T., Ratnakar, V., Gupta, S., Muslea, M., Silva, F., & Knoblock, C. (2011). Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows. *10th International Semantic Web Conference*, Bonn (Germany) (pp. 65–80).

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone?. *First Monday*, 16(2).

Heald, D. (2012). Why is transparency about public expenditure so elusive?. *International Review of Administrative Sciences*, 1, 30–49.

Hernandez-Perez, T., Rodriguez-Mateos, D., Martin-Galan, B, & Antonia Garcia-Moreno, M. (2009). Use of metadata in Spanish electronic e-government: the challenges of interoperability. *Revista Española de Documentacion Cientifica*, 32(4), 64–91.

Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, 12(1), 4–16.

IANA. (1988). Retrieve from http://iana.org (accessed on Nov. 2017).

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, 28(4), 446–456.

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.

Jarrar, Y., Schiuma, G., & Salem, F. (2007). Benchmarking the e-government bulldozer: Beyond measuring the tread marks. *Measuring business excellence*, 11(4), 9–22.

Koussouris, S., Lampathaki, F., Kokkinakos, P., Askounis, D., & Misuraca, G. (2015). Accelerating Policy Making 2.0: Innovation directions and research perspectives as distilled from four standout cases. *Government Information Quarterly*, 32(2), 142–153.

Kubler, S., Robert, J., Derigent, W., Voisin, A., & Le Traon, Y. (2016a). A state-of the-art survey & testbed of Fuzzy AHP (FAHP) applications. *Expert Systems with Applications*, 65, 398–422.

Kubler, S., Robert, J., Le Traon, Y., Umbrich, J., & Neumaier, S. (2016b). Open Data Portal Quality Comparison Using AHP. *17th International Digital Government Research Conference on Digital Government Research*, Shanghai (China) (pp. 397–407).

Kučera, J., Chlapek, D., & Nečaskỳ, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. *International Conference on Electronic Government and the Information Systems Perspective*, Prague (Czech Republic) (pp. 152–166).

van Laarhoven, P.J.M., & Pedrycz, W. (1983). A fuzzy extension of Saaty's priority theory. *Fuzzy Sets and Systems*, 11(1), 199–227.

Lee, G., & Kwak, Y. H. (2012). An open government maturity model for social media-based public engagement. *Government Information Quarterly*, 29(4), 492–503.

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, 32(3), 323–332.

Mardani, A., Jusoh, A., & Zavadskas, E. K. (2015). Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014. *Expert Systems with Applications*, 42(8), 4126–4148.

Martin, S., Foulonneau, M., & Turki, S. (2013). 1-5 stars: Metadata on the openness level of open data sets in Europe. *Research Conference on Metadata and Semantic Research*, Karlsruhe (Germany) (pp. 234–245).

Moreno-Jiménez, J. M., Pérez-Espés, C. & Velázquez, M. (2014). e-Cognocracy and the design of public policies. *Government Information Quarterly*, 31(1), 185–194.

Neumaier, S., & Umbrich, J. (2016). Measures for assessing the data freshness in Open Data portals. *International Conference on Open and Big Data*, Vienna (Austria), (pp. 17–24).

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information quality*, 8(1), 2:1–2:29.

Open Data Institute. Benchmarking open data automatically. (2015). Retrieve from https://theodi.org/guides/benchmarking-data-automatically (accessed on Nov. 2017).

Open Government Partnership. (2011). Retrieve from https://www.opengovpartnership.org (accessed on Nov. 2017).

Open Government Working Group. 8 Principles of Open Government Data. (2007). Retrieve from https://public.resource.org/8_principles.html (accessed on Nov. 2017).

Open Knowledge International. Open Definition 2.1. (2017). Retrieve from http://opendefinition.org/od/ (accessed on Nov. 2017).

Ouzzani, M., Papotti, P., & Rahm, E. (2013). Introduction to the special issue on data quality. *Information Systems*, 38(6), 885–886.

Parycek, P., & Sachs, M. (2010). Open government–information flow in Web 2.0. *European Journal of ePractice*, 9(1), 1–70.

Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. *Conference for E-Democracy and Open Governement*, Krems an der Donau (Austria) (335).

Ren, G.-J., & Glissmann, S (2012). Identifying information assets for open data: the role of business architecture and information quality. *IEEE 14th International Conference on Commerce and Enterprise Computing*, Hangzhou (China) (pp. 94–100).

Rojas, L. A. R., Bermdez, G. M. T., & Lovelle, J. M. C. (2014). Open Data and Big Data: A Perspective from Colombia. In Uden, L., Fuenzaliza Oshee, D., Ting, I.-H., & Liberona, D. (Eds.) *The 8th International Conference on Knowledge Management in Organizations* (pp. 35–41). Springer International Publishing.

Saaty, T. L. (1986). Absolute and relative measurement with the AHP. The most livable cities in the United States. *Socio-Economic Planning Sciences*, 20(6), 327–331.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, 15(3), 234–281.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly*, 32(3), 308–315.

Sugimoto, S. (2014). Digital archives and metadata as critical infrastructure to keep community memory safe for the future–lessons from Japanese activities. *Archives and Manuscripts*, 42(1), 61–72.

Tolbert, C. and Mossberger, K. (2006). The effects of e-government on trust and confidence in government. *Public Administration Review*, 66(3), 354–369.

Tygel, A., Auer, S., Debattista, J., Orlandi, F., & Campos, M. L. M. (2016). Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach. *IEEE 10th International Conference onSemantic Computing*, Laguna Hills (USA) (pp. 71–78).

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, 31(2), 278–290.

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.

W3C, Data Catalog Vocabulary (DCAT). (2014). Retrieve from http://www.w3.org/TR/vocab-dcat/ (accessed on Nov. 2017).

Waugh, P. Improving data quality on data.gov.au. (2015). Retrieve from https://blog.data.gov.au/news-media/blog/improving-data-quality-datagovau (accessed on Nov. 2017).

Welle Donker, F., & van Loenen, B. (2017). How to assess the success of the open data ecosystem?. *International Journal of Digital Earth*, 10(3), 284–306.

Zissis, D., & Lekkas, D. (2011). Securing e-Government and e-Voting with an open cloud computing architecture. *Government Information Quarterly*, 28(2), 239–251.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical impediments of open data. *Electronic Journal of eGovernment*, 10(2), 156–172.

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012). The potential of metadata for linked open data and its value for users and publishers. *JeDEM-eJournal of eDemocracy and Open Government*, 4(2), 222–244.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1), 17–29.

Zuiderwijk, A., & Janssen, M. (2014). The negative effects of open government data - investigating the dark side of open data. *15th Annual International Conference on Digital Government Research*, Phoenix (USA) (pp. 147–152).

Zuiderwijk, A., Janssen, M., Susha, I. (2016). Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 116–146.