Open Data Portal Quality Comparison using AHP

Sylvain Kubler
University of Luxembourg
Interdisciplinary Centre for
Security, Reliability & Trust
4 rue Alphonse Weicker
L-2721 Luxembourg
sylvain.kubler@uni.lu

Jérémy Robert
University of Luxembourg
Interdisciplinary Centre for
Security, Reliability & Trust
4 rue Alphonse Weicker
L-2721 Luxembourg
jeremy.robert@uni.lu

Yves Le Traon
University of Luxembourg
Interdisciplinary Centre for
Security, Reliability & Trust
4 rue Alphonse Weicker
L-2721 Luxembourg
yves.letraon@uni.lu

Jürgen Umbrich
Vienna University of
Economics and Business,
Institute for Information
Business, Welthandelsplatz 1
1020 Vienna, Austria
juergen.umbrich@wu.ac.at

Sebastian Neumaier
Vienna University of
Economics and Business,
Institute for Information
Business, Welthandelsplatz 1
1020 Vienna, Austria
sebastian.neumaier@wu.ac.at

ABSTRACT

During recent years, more and more Open Data becomes available and used as part of the Open Data movement. However, there are reported issues with the quality of the metadata in data portals and the data itself. This is a serious risk that could disrupt the Open Data project, as well as *e*-government initiatives since the data quality needs to be managed to guarantee the reliability of *e*-government to the public. First quality assessment frameworks emerge to evaluate the quality for a given dataset or portal along various dimensions (e.g., information completeness). Nonetheless, a common problem with such frameworks is to provide meaningful ranking mechanisms that are able to integrate several quality dimensions and user preferences (e.g., a portal provider is likely to have different quality preferences than a portal consumer). To address this multi-criteria decision making problem, our research work applies AHP (Analytic Hierarchy Process), which compares 146 active Open Data portals across 44 countries, powered by the CKAN software.

Keywords

Open Data; *e*-Government; Data Quality; Multi-Criteria Decision Making; Analytic Hierarchy Process

1. INTRODUCTION

The concept of Open Data, which is data published under a license that allows everybody to (re-)use and modify the content, gains importance in the context of a growing demand for transparency in the public and private sector. Organizations from all over the world are under increasing pressure to release their data to a variety of users (citizens, businesses, academics. . .), leading to increased public transparency and allowing for enhanced data-enriched public engagement in policy and other analysis [9]. Additionally, it is expected that Open Data supports the decision making of both governments and individuals [12, 6]. In particular, the public is expected to be able to use government data to make better decision and improve the quality of their lives (e.g., by making specific databases easily accessible through mobile apps) [5, 14], while governments are expected to be able to more easily access a wider range of datasets to foster evidence-based decision making [1].

Although opportunities are wide and worth exploring, first critical voices and reports confirmed that there exists a quality problem in Open Data [12, 16, 21]. This is a serious risk that could disrupt the Open Data project since data consumers normally search over the metadata to discover, locate and use relevant data sources in their own core businesses. Missing or incorrect information in the metadata has a non-negligible impact on *i)* open and *e*-government since the quality of the published data is one of the key factors to be taken into account in the *e*-government assessment process (e.g., to validate whether *e*-government goals are or not satisfied) [10], as well as on *ii)* businesses since data quality is a key factor that can determine whether these metadata can be useful for businesses or not (if the quality is not high, businesses would like know about it) [21]. The information quality (or data quality[1]) of Open Data if often said to be mixed, i.e. depending upon several quality dimensions. In this regard, there is still a lack of frameworks and tools to dynamically assess the data quality in place [20] and compare Open Data portals with one another based upon such distinct dimensions, where quality can be high on one dimension, but low on another one. Add to that the fact that Open Data users may have, according to their needs and core businesses, different preferences regarding one or more quality dimensions (e.g., openness might be seen as more important than completeness, and *vice-versa*).

dg.o '16, June 08-10, 2016, Shanghai, China

DOI: http://dx.doi.org/10.1145/2912160.2912167

[1]"Information" is often described as "data" that has been processed in some manner, but this article uses both terms interchangeably.

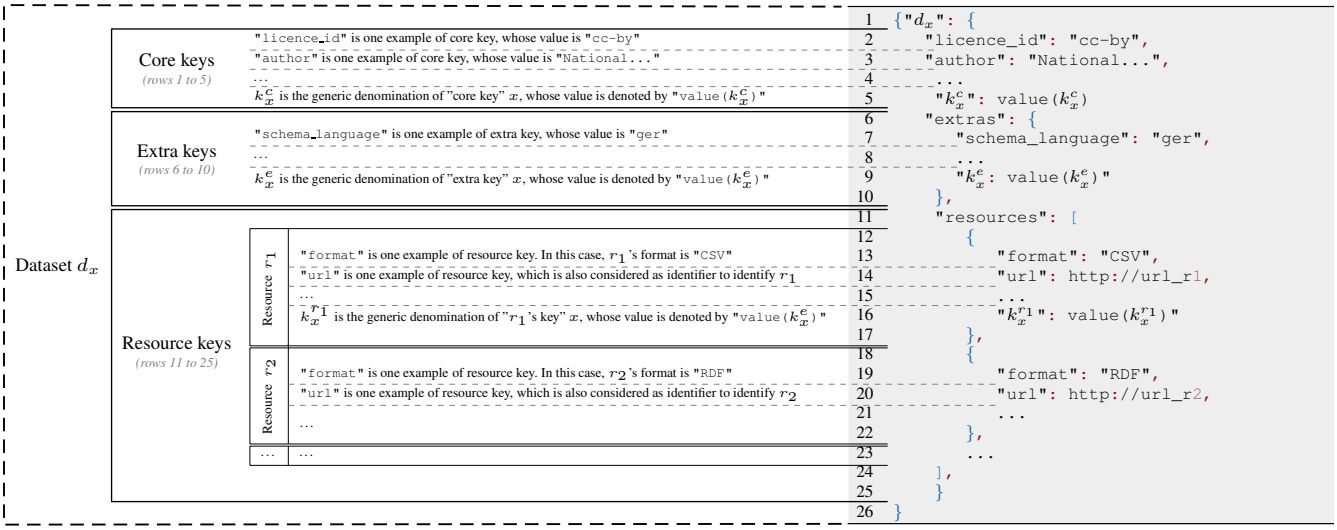


Figure 1: High level structure of the meta data for a CKAN dataset

To date, there is no framework based on well-defined and transparent MCDM (multi-criteria decision making) assessment formulas to compare open data portal quality (OPDQ) with regard to specific user preferences. This paper addresses this issue by introducing a methodology based on Analytic Hierarchy Process (AHP). This methodology is further turned into an ODPQ Web dashboard enabling any Open Data end-user to identify, at any point in time, the quality and ranking of one or a group of Open Data portals. Section 2 provides an overview of the most widely used platforms for the publication and management of Open Data, discusses a set of quality dimensions related to Open Data portals, and introduces the MCDM problem for aggregating all those quality dimensions. Section 3 presents how this problem is tackled using AHP. Section 4 presents the resulting ODPQ Web dashboard, which is currently monitoring, assessing and comparing a set of over 146 active Open Data portals; discussion and conclusion follow.

2. OPEN DATA PORTALS & QUALITY DIMENSIONS

Open Data portals can be seen as digital catalogues containing dataset descriptions. Those dataset description typically consists of a set of key value pairs (commonly referred to as metadata) to describe important contextual information such as license information, authorship, timeliness, or data formats about an actual data source. Section 2.1 gives insight into existing platforms for publishing Open Data, along with some formal definitions used in the rest of this paper. In section 2.2, we recapture quality dimensions and associated metrics from previous work. Finally, section 2.3 discusses the MCDM nature of the Open Data portal comparison problem, along with the research methodology.

2.1 Open Data Platforms

To accelerate the usage of data by citizens and developers, it is necessary to adopt an effective Open Data publishing ecosystem. Such an Open Data platform serves as a single point of access to advertise and consume datasets, includes a human user interface and potentially APIs for legal agents. Examples of such platforms are the commercial *Socrata* Open Data portal software[2], the community-based *CKAN*[3] software, or still *OpenDataSoft*[4] that is mainly used for French Open Data catalogs.

The Socrata software, funded in 2007, provides a cloud-based service for data publishing, metadata management, data catalogue federation, and exposure of data (data can be published manually, or through dedicated APIs). CKAN (Comprehensive Knowledge Archive Network) is an open-source platform maintained by the Open Knowledge Foundation. In contrast to Socrata, CKAN allows both, the upload and storage of the data on the portal server or references to external data sources. Out of the presented Open Data publishing platforms, CKAN is the most popular framework and, for instance, is used by the US, UK, and Canadian government, as well as by the recently issued European Open Data portal[5] (developed by the Belgian company Tenforce).

In previous research, we focused on monitoring and assessing CKAN portals. The central entities in any CKAN portal are datasets, which contain general metadata to describe important contextual information about the dataset and its data sources. CKAN's dataset metadata is natively published in JSON format, as depicted in Figure 1 (simplified view of a dataset denoted by d_x). Further, we distinguish three categories of meta data keys in a CKAN portal:

- *core keys:* a set of predefined keys that are generic and restrictive, and by default available in any CKAN portal (e.g., `license_ID` as shown in Figure 1, row 2);
- *extra keys:* a set of arbitrary additional meta data keys to describe a datasets defined by the portal provider. These keys are listed under the `extra` key (*cf.* Figure 1 – *cf.* rows 6-10);
- *resource keys:* a mix between some default keys and additional keys defined by the portal provider to describe the particular resources (e.g., a datafile, API. . .).

[2] https://opendata.socrata.com
[3] http://ckan.org
[4] https://www.opendatasoft.com
[5] http://open-data.europa.eu/

Table 1: Quality dimensions impacting on the Open Data Portal quality

Dimensions	Sub-dimensions		Description	Metric
Usage (Q_u)	Core	$Q_{u(core)}$	The extent to which available meta data 'core' keys are used to describe a dataset	%
	Resource	$Q_{u(res)}$	The extent to which available meta data 'resource' keys are used to describe a dataset	%
	Extra	$Q_{u(extra)}$	The extent to which available meta data 'extra' keys are used to describe a dataset	%
Completeness (Q_c)	Core	$Q_{c(core)}$	The extent to which the used meta data 'core' keys are non empty	%
	Resource	$Q_{c(res)}$	The extent to which the used meta data 'resource' keys are non empty	%
	Extra	$Q_{c(extra)}$	The extent to which the used meta data 'extra' keys are non empty	%
Openness (Q_o)	File	$Q_{o(file)}$	The extent to which 'license' formats conform to the open definition (*cf.* section 2.2)	%
	Licence	$Q_{o(lic)}$	The extent to which 'file' formats conform to the open definition (*cf.* section 2.2)	%
Addressability (Q_a)	URL	$Q_{a(url)}$	The extent to which the data publisher provides contact information via 'URL'	%
	Email	$Q_{a(email)}$	The extent to which the data publisher provides contact information via 'email'	%
Retrievability (Q_{ds})	Dataset	$Q_{a(ds)}$	The extent to which the dataset can be retrieved without any error or access restriction	%
	Resource	$Q_{a(res)}$	The extent to which the dataset can be retrieved without any error or access restriction	%

Each resource is described under the `resources` key, as emphasized in Figure 1 (*cf.* rows 12-17; 18-22).

A set of quality dimensions, based on each of the above categories, have been defined in our previous work [19], as will be discussed in the next section.

2.2 Open Data Quality Dimensions

Several quality dimensions have been introduced in the literature related to open data (see e.g. [3, 2]). Our study considers some of those dimensions, and particularly the ones introduced in [19], namely:

- *Usage (Q_u):* The usage metric, formally detailed in [19], is the degree (%) to which the available meta data keys are used in the datasets of a given portal. We use this measure since we observed that not all portals make all meta data keys available to the data publishers or because keys can be left out if publishers use the CKAN API. While this usage metric is a rather weak quality measure, it can be used either as a weight for other quality formulas or as a filter, e.g., one can compute a certain metric by considering only the keys that are used in all datasets (*cf.* Figure 1);

- *Completeness (Q_c):* The completeness of the meta data description is a widely used and important measure to provide an indication of how much meta information is available for a given dataset. Formally, the completeness of a portal is the degree (%) to which the available meta data keys to describe a dataset have non empty values;

- *Openness (Q_o):* The openness of a portal is the degree (%) to which datasets provide a confirmed open license and to which the resources are available in an appropriate Open Data file format (wrt. the Open Definition[6]). Although a semantic distance could be introduced in order to highlight how "open" a license or format is (e.g., usage of the data is allowed but not the redistribution of modified values), it is very challenging to define such a distance for all licenses and formats. Accordingly, the metric proposed in [19] evaluates whether the specified license (per dataset) is included in the list provided by the Open Definition. This list contains details about 108 different licenses including their typical `id`, `url`, `title` and an assessment whether or not they are considered as "open". Regarding the "format" openness metric, it is worth noting that a dataset can have various resources with different formats (CSV, RDF, HTML, LaTeX, XML, JSON, TXT, GIF...), as depicted in Figure 1 with r_1 and r_2. A dataset is labelled as open as soon as one resource of the dataset has an open format;

- *Addressability (Q_a):* Addressability is another important dimension of Open Data portals since it emphasizes the extent to which contact information about the dataset's creator/maintainer is made available. Formally, the proposed metric defines the degree (%) to which datasets provide a value, an email address or HTTP URL to contact the data publisher [19];

- *Retrievability (Q_r):* Retrievability emphasizes the extent to which resource and datasets are freely accessible as a whole. More concretely, it measures whether a legal or software agent can retrieve the content of a portal and its resources (based on an `HTTP GET` operation) without any error or access restriction.

These five dimensions are partially aligned with existing ones [16], and extended by Openness and Addressability. Each dimension has been broken down into sub-dimensions, as summarized in Table 1.

2.3 Open Quality Assessment Methodology

A simplistic view of the overall portal quality assessment metholodoly is given in Figure 2, which starts by *(i)* crawling and collecting datasets from distinct Open Data (CKAN) portals (see Stage 1), then *(ii)* assessing each dataset based on the introduced quality dimensions (see Stage 2, where 'smileys' illustrate whether portals 1 and 2 are positively or negatively assessed with respect to each sub-dimension); and finally *(iii)* aggregating all quality results, as well as the user requirements/preferences (e.g., prioritization of one or more quality dimensions) in order to obtain a final ranking of the Open Data portals' quality. Our previous research [19] focused on stages 1 and 2 (see Figure 2), but so far the MCDM nature of the comparison process has not yet been addressed.

Over the last three decades, a number of MCDM techniques have been developed such as AHP (analytic hierarchy process), ANP (analytic network process), ELECTRE, *etc.* [13, 7]. There are no better or worse techniques but some techniques are better suited to particular decision-making problems than others. In our study, AHP is used for two main reasons: *i)* we only deal with linear preferences and

[6] http://opendefinition.org/

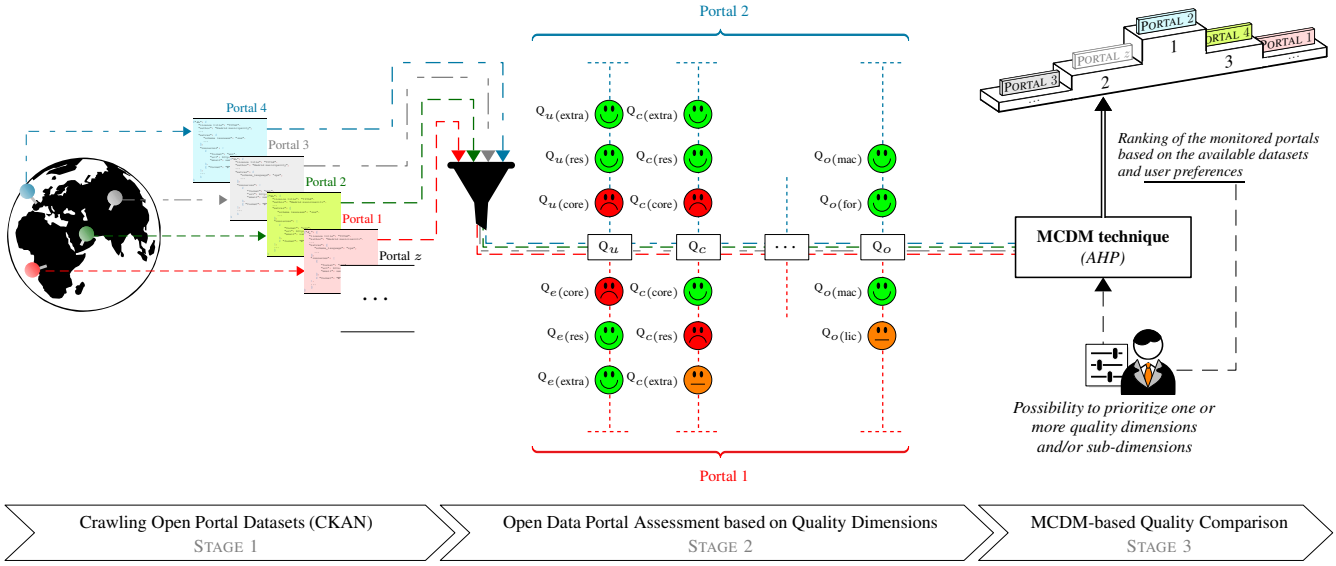


Figure 2: Overall quality assessment methodology: from metadata collection to Open Data portals' ranking

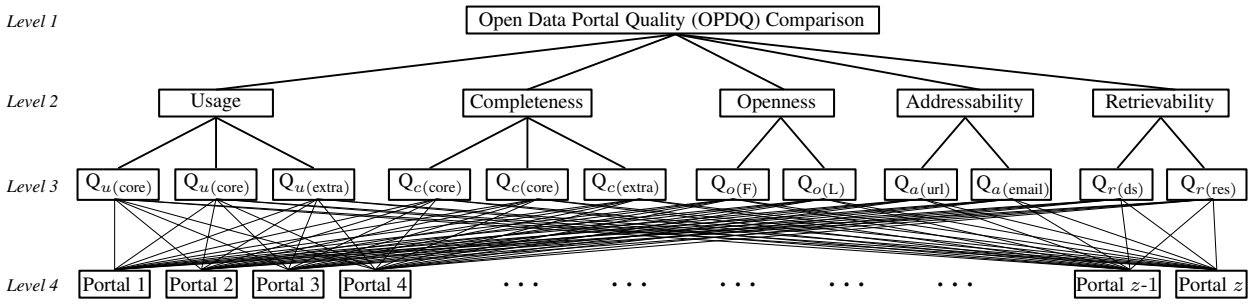


Figure 3: AHP structure of the Open Data Portal quality comparison process

ii) AHP enables to easily and effectively integrate[7] expert requirements/preferences as well as tangible system parameters/characteristics. The next section focuses on stage 3 and highlights how AHP is applied to our problem.

3. AHP-BASED COMPARISON

AHP, originally introduced by [18], 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, namely: the overall goal, the criteria (potential sub-criteria) and the alternatives. The MCDM ranking problem of our study is broken down into the hierarchical structure depicted in Figure 3, which consists of four distinct levels, namely *(i) Level 1:* the overall goal of the study is to assess and rank the monitored Open Data portals in terms of published metadata quality; *(ii) Levels 2 and 3:* the set of quality dimensions and sub-criteria introduced in Table 1; *(iii) Level 4* the monitored Open Data portals that correspond to the alternatives.

Given this hierarchy, AHP does perform the following computation steps for identifying the final ranking of the alternatives with respect to the overall goal:

1. Compare each element in the corresponding level and calibrate them on the numerical scale. This requires $\frac{n(n-1)}{2}$ pairwise comparisons, where n is the number of elements (diagonal elements being equal to "1" and the other elements being the reciprocal of the earlier comparisons);
2. Perform calculation to find the maximum eigenvalue, consistency index (CI), consistency ratio (CR), and normalized values;
3. If CI and CR are satisfactory, then decision/ranking is done based on the normalized eigenvalues.

These three stages are detailed in the following sections, in which a scenario – *whose parts are preceded by the symbol "↪"* – is considered to make the understanding easier.

3.1 Pairwise comparison based preference measurement

According to [4], two types of judgment exist: *"Comparative judgment* and *"Absolute judgment.* In comparative/relative measurement, each alternative is compared with many other alternatives, that is why this is also referred to as "pairwise comparisons as ratios" in the AHP literature

[7] According to a recent survey on MCDM techniques [13], AHP is the second most used technique with a frequency of application of 15.82%, followed by Fuzzy AHP (9.53%).

Table 2: Variable definitions

Variables	Description
Q_x	abbreviation for Quality dimension x with $x = \{1, 2, .., m\}$. In this study, five dimensions are defined at level 2 of the hierarchy structure, namely: Q_u, Q_c, Q_o, Q_a and Q_r, as defined in *cf.* Table 1.
$\mathrm{Q}_{x(h)}$	abbreviation for a sub-dimension of dimension x; e.g. $h = \{\mathrm{Q}_{u(\mathrm{core})}, \mathrm{Q}_{u(\mathrm{res})}, \mathrm{Q}_{u(\mathrm{extra})}\}$ for $x = u$, as summarized in Table 1.
P_c	abbreviation for ""Pairwise Comparison matrix", whether at level 2, 3 or 4 of the AHP structure.
w_{ij}	crisp value of a pairwise comparison matrix located at row i, column j of P_c.
A_l	alternative $l = \{1, 2, .., z\}$ in the AHP structure with z the number of monitored Open Data portals.
W_{Q_x}, $W_{\mathrm{Q}_{x(h)}}$	eigenvalue of dimension Q_x or sub-dimension $\mathrm{Q}_{x(h)}$ (the eigenvalue being computed from P_c). In practice, it indicates the importance of one (sub-)dimension against the others.
$\mathrm{M}^{\mathrm{A}_l}_{\mathrm{Q}_{x(h)}}$	measurable metric (expressed as a &, as given in Table 1) assessing portal A_l with respect to the quality dimension $\mathrm{Q}_{u(h)}$.
$W^{\mathrm{A}_l}_{\mathrm{Q}_{x(h)}}$	eigenvalue of alternative A_l with respect to sub-dimension $\mathrm{Q}_{x(h)}$. In practice, it indicates how good (or bad) the quality of portal l is with respect to $\mathrm{Q}_{x(h)}$.

[15]. In absolute measurement, each alternative is compared with an ideal alternative the expert knows of or can imagine, that is why this is referred to as "pairwise comparison based preference measurement". This section details the "pairwise comparison based preference measurement" principle that is applied at level 2 and 3 of the AHP structure (*cf.* Figure 3), while section 3.2 details the "pairwise comparisons as ratios" applied at level 4. Note that all variables used in this paper are summarized in Table 2.

In pairwise comparison-based preference measurement, decision makers evaluate the importance of one dimension with respect to the others. Pairwise comparisons among quality dimensions are performed at a same level, as formalized in Eq. 1, with m the number of dimensions to be compared (e.g., at level 2 of AHP, $m = |\{\mathrm{Q}_u, \mathrm{Q}_c, \mathrm{Q}_o, \mathrm{Q}_a, \mathrm{Q}_r\}| = 5$). The evaluation carried out by the expert is based on the 1- to 9-point Saaty's scale: $\{1, 3, 5, 7, 9\}$; $w_{ij} = 1$ meaning that Q_i and Q_j are of equal importance and $w_{ij} = 9$ meaning that Q_i is strongly favored over Q_j. The computation of the normalized eigenvector of P_C enables to turn qualitative data into crisp ratios [17]. Although not detailed in this paper, it is important to note that a consistency ratio (CR) has to be computed to identify whether the end-user's answers are or not consistent; a pairwise comparison is regarded as acceptable if CR< 10% [17].

$$Pc = \begin{array}{c} \\ \mathrm{Q}_1 \\ \vdots \\ \mathrm{Q}_m \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q}_1 & \dots & \mathrm{Q}_m \end{array} \\ \begin{bmatrix} w_{11} & \dots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \dots & w_{mm} \end{bmatrix} \end{array} \quad (1)$$

➥ Eq. 2 shows the user preference specifications related to the quality dimensions defined at Level 2 of the AHP structure. The computed normalized eigenvector highlights that the end-user respectively prioritizes Usage (Q_u), Completeness (Q_c) and Openness (Q_o) over the Addressability (Q_a) and Retrievability (Q_r) dimensions (see $\mathrm{W}_{\mathrm{Q}_x}$ in Eq. 2).

$$\begin{array}{c} \\ \mathrm{Q}_u \\ \mathrm{Q}_c \\ \mathrm{Q}_o \\ \mathrm{Q}_a \\ \mathrm{Q}_r \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{Q}_u & \mathrm{Q}_c & \mathrm{Q}_o & \mathrm{Q}_a & \mathrm{Q}_r \end{array} \\ \begin{bmatrix} 1 & 3 & 1 & 3 & 9 \\ 1/3 & 1 & 1 & 5 & 5 \\ 1 & 1 & 1 & 3 & 3 \\ 1/3 & 1/5 & 1/3 & 1 & 1 \\ 1/9 & 1/5 & 1/3 & 1 & 1 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ W_{\mathrm{Q}_u} \\ W_{\mathrm{Q}_c} \\ W_{\mathrm{Q}_o} \\ W_{\mathrm{Q}_a} \\ W_{\mathrm{Q}_r} \end{array} \begin{bmatrix} 0.39 \\ 0.24 \\ 0.24 \\ 0.07 \\ 0.06 \end{bmatrix} \quad (2)$$

Eq. 3 shows an example of pairwise comparisons carried out at Level 3 of the AHP structure, considering the sub-dimensions of Q_o, namely $\{\mathrm{Q}_{o(\mathrm{F})}, \mathrm{Q}_{o(\mathrm{L})}\}$. The resulting eigenvector (see Eq. 3) shows that the end-user puts a higher priority on the openness of the "Format" of datasets ($\mathrm{Q}_{o(\mathrm{F})}$) than on the "Licence" aspect ($\mathrm{Q}_{o(\mathrm{L})}$).

$$\begin{array}{c} \\ \mathrm{Q}_{o(\mathrm{F})} \\ \mathrm{Q}_{o(\mathrm{L})} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{Q}_{o(\mathrm{F})} & \mathrm{Q}_{o(\mathrm{L})} \end{array} \\ \begin{bmatrix} 1 & 5 \\ \frac{1}{5} & 1 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ W_{\mathrm{Q}_{o(\mathrm{F})}} \\ W_{\mathrm{Q}_{o(\mathrm{L})}} \end{array} \begin{bmatrix} 0.83 \\ 0.17 \end{bmatrix} \quad (3)$$

Although all the numerical examples are not detailed here, it must be noted that all the other pairwise comparisons at level 2 and 3 must similarly be carried out by the end-user.

3.2 Pairwise comparisons as ratio measurement

Pairwise comparisons as ratios are applied at level 4 of the AHP structure in order to compare – *based upon the measurable metrics given in Table 1* – alternatives with each other, and with respect to each quality dimension. To this end, Eq. 4 gives insight into such a ratio matrix, where $\mathrm{M}^{\mathrm{A}_l}_{\mathrm{Q}_{x(h)}}$ corresponds to the metric assessing portal A_l with respect to sub-dimension $\mathrm{Q}_{x(h)}$. The normalized eigenvector values with respect to $\mathrm{Q}_{u(h)}$ are denoted by $W^{\mathrm{A}_l}_{\mathrm{Q}_{u(h)}}$.

$$\begin{array}{c} \\ \mathrm{A}_1 \\ \mathrm{A}_2 \\ \vdots \\ \mathrm{A}_z \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A}_1 & \mathrm{A}_2 & \dots & \mathrm{A}_z \end{array} \\ \begin{bmatrix} 1 & \frac{\mathrm{M}^{\mathrm{A}_1}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_2}_{\mathrm{Q}_{x(h)}}} & \dots & \frac{\mathrm{M}^{\mathrm{A}_1}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_z}_{\mathrm{Q}_{x(h)}}} \\ \frac{\mathrm{M}^{\mathrm{A}_2}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_1}_{\mathrm{Q}_{x(h)}}} & 1 & \dots & \frac{\mathrm{M}^{\mathrm{A}_2}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_z}_{\mathrm{Q}_{x(h)}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{M}^{\mathrm{A}_z}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_1}_{\mathrm{Q}_{x(h)}}} & \frac{\mathrm{M}^{\mathrm{A}_z}_{\mathrm{Q}_{x(h)}}}{\mathrm{M}^{\mathrm{A}_2}_{\mathrm{Q}_{x(h)}}} & \dots & 1 \end{bmatrix} \end{array} \Rightarrow \begin{bmatrix} W^{\mathrm{A}_1}_{\mathrm{Q}_{x(h)}} \\ W^{\mathrm{A}_2}_{\mathrm{Q}_{x(h)}} \\ \vdots \\ W^{\mathrm{A}_z}_{\mathrm{Q}_{x(h)}} \end{bmatrix} \quad (4)$$

➥ Let us consider pairwise comparisons as ratios between Portal 1 and 2 with respect to $\mathrm{Q}_{o(\mathrm{F})}$. Portal 1 has 1000 available datasets, 437 of them (i.e., 43.7% – see Eq. 5) are 'open' according to the definition given in section 2.2, while openness reaches 66.2% on Portal 2 (see Eq. 5). The resulting pairwise comparisons as ratio matrix with respect to $\mathrm{Q}_{o(\mathrm{F})}$ is given in Eq. 6, where the two metrics computed in Eq. 5 are used in row 1/column 2 of the matrix, and *vice-versa*. The resulting eingevector (set of $W^{\mathrm{A}_l}_{\mathrm{Q}_{o(F)}}$ in this example) thus indicates how good/bad the quality of each portal is, with respect to the considered quality dimension.

$$\mathrm{M}^{\mathrm{A}_1}_{\mathrm{Q}_{o(\mathrm{F})}} = \frac{437}{1000} = 43.7\% \quad \mathrm{M}^{\mathrm{A}_2}_{\mathrm{Q}_{o(\mathrm{F})}} = \frac{2443}{3690} = 66.2\% \quad (5)$$

$$\begin{array}{c} \\ \mathrm{A}_1 \\ \mathrm{A}_2 \\ \vdots \\ \mathrm{A}_z \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A}_1 & \mathrm{A}_2 & \dots & \mathrm{A}_z \end{array} \\ \begin{bmatrix} 1 & \frac{\mathbf{43.7}}{\mathbf{66.2}} & \dots & 1.397 \\ \frac{\mathbf{66.2}}{\mathbf{43.7}} & 1 & \dots & 2.115 \\ \vdots & \vdots & \ddots & \vdots \\ 0.716 & 0.473 & \dots & 1 \end{bmatrix} \end{array} \rightarrow \begin{array}{c} W^{\mathrm{A}_1}_{\mathrm{Q}_{o(\mathrm{F})}} \\ W^{\mathrm{A}_2}_{\mathrm{Q}_{o(\mathrm{F})}} \\ \vdots \\ W^{\mathrm{A}_z}_{\mathrm{Q}_{o(\mathrm{F})}} \end{array} \begin{bmatrix} 0.0135 \\ 0.0097 \\ \vdots \\ 0.0010 \end{bmatrix} \quad (6)$$

Section 3.3 presents how the different AHP scores are aggregated in order to obtain the final quality ranking of the monitored portals.

3.3 TOPSIS-based alternative ranking

The set of scores computed in the previous sections are then turned into a global weight based on Eq. 7, considering each alternative A_l with respect to each sub-dimension $\mathrm{Q}_{x(h)}$, and their respective parent (i.e., Q_x). All those global weights are summarized in the form of a matrix in Eq. 8.

$$\mathring{W}^{\mathrm{A}_l}_{\mathrm{Q}_{x(h)}} = W^{\mathrm{A}_l}_{\mathrm{Q}_{x(h)}} \times W_{\mathrm{Q}_{x(h)}} \times W_{\mathrm{Q}_x} \quad (7)$$

$$\begin{array}{c} \\ \mathrm{A}_1 \\ \mathrm{A}_2 \\ \vdots \\ \mathrm{A}_z \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{Q}_{u(\mathrm{core})} & \dots & \mathrm{Q}_{u(\mathrm{extra})} & \dots & \mathrm{Q}_{r(\mathrm{res})} \end{array} \\ \begin{bmatrix} \mathring{W}^{\mathrm{A}_1}_{\mathrm{Q}_{u(\mathrm{core})}} & \dots & \mathring{W}^{\mathrm{A}_1}_{\mathrm{Q}_{u(\mathrm{extra})}} & \dots & \mathring{W}^{\mathrm{A}_1}_{r(\mathrm{res})} \\ \mathring{W}^{\mathrm{A}_2}_{\mathrm{Q}_{u(\mathrm{core})}} & \dots & \mathring{W}^{\mathrm{A}_2}_{\mathrm{Q}_{u(\mathrm{extra})}} & \dots & \mathring{W}^{\mathrm{A}_2}_{r(\mathrm{res})} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathring{W}^{\mathrm{A}_z}_{\mathrm{Q}_{u(\mathrm{core})}} & \dots & \mathring{W}^{\mathrm{A}_z}_{\mathrm{Q}_{u(\mathrm{extra})}} & \dots & \mathring{W}^{\mathrm{A}_z}_{r(\mathrm{res})} \end{bmatrix} \end{array} \quad (8)$$

➥ For illustration purposes, Eq. 9 details the global weight calculation for A_1 (i.e., Portal 1) with respect to the sub-dimension $\mathrm{Q}_{o(\mathrm{F})}$, and its respective parent Q_o.

$$\begin{aligned} \mathring{W}^{\mathrm{A}_1}_{\mathrm{Q}_{o(\mathrm{F})}} &= W^{\mathrm{A}_1}_{\mathrm{Q}_{o(\mathrm{F})}} \times W_{\mathrm{Q}_{o(\mathrm{F})}} \times W_{\mathrm{Q}_o} \\ &= 0.0135 \times 0.83 \times 0.24 = 0.00269 \end{aligned} \quad (9)$$

Finally, the global weights can be aggregated to obtain the final quality score, based on which the final portal ranking is generated. Although a variety of aggregation methods exists in the literature (see e.g. [7]), our study uses TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) to generate the final quality scores and alternative ranking. Technically, TOPSIS introduces for each alternative A_l the closeness coefficient denoted by $R(\mathrm{A}_l)$. To compute this coefficient, the positive ideal solution (PIS) denoted by $d^+_{\mathrm{Q}_{x(h)}}$, and negative ideal solution (NIS) denoted by $d^-_{\mathrm{Q}_{x(h)}}$, are computed for each sub-dimension $\mathrm{Q}_{x(h)}$ as formalized in Eq. 10. The distances measuring the separation from PIS and NIS, respectively denoted by $D^+_{\mathrm{A}_l}$ and $D^-_{\mathrm{A}_l}$, are then computed in Eq. 11 and 12.

$$d^+_{\mathrm{Q}_{x(h)}} = \underset{l=1..z}{max}\left(\mathring{W}^{\mathrm{A}_l}_{\mathrm{Q}_{xh}}\right) \qquad d^-_{\mathrm{Q}_{x(h)}} = \underset{l=1..z}{min}\left(\mathring{W}^{\mathrm{A}_l}_{\mathrm{Q}_{xh}}\right) \quad (10)$$

$$D^+(\mathrm{A}_l) = \sqrt{\sum_{xh}\left(\mathring{W}^{\mathrm{A}_l}_{\mathrm{Q}_{xh}} - d^+_{\mathrm{Q}_{x(h)}}\right)^2} \quad l = 1, .., z \quad (11)$$

$$D^-(\mathrm{A}_l) = \sqrt{\sum_{xh}\left(\mathring{W}^{\mathrm{A}_l}_{\mathrm{Q}_{xh}} - d^-_{\mathrm{Q}_{x(h)}}\right)^2} \quad l = 1, .., z \quad (12)$$

$$R(\mathrm{A}_l) = \frac{D^-(\mathrm{A}_l)}{D^+(\mathrm{A}_l) + D^-(\mathrm{A}_l)} \quad l = 1, .., z \quad (13)$$

Table 3: Alternative ranking illustration

	Ranking per quality dimension					Final
	Q_u	Q_c	Q_o	Q_a	Q_r	
Portal 1	70^{th}	68^{th}	63^{th}	55^{rd}	25^{rd}	43^{th}
Portal 2	106^{th}	55^{th}	115^{th}	87^{th}	27^{rd}	85^{th}
...	...	...	...	...	...	...
Portal 17	$\mathbf{111}^{st}$	$\mathbf{100}^{rd}$	$\mathbf{105}^{th}$	$\mathbf{108}^{nd}$	$\mathbf{123}^{th}$	$\mathbf{121}^{rd}$
...	...	...	...	...	...	...
Portal 41	$\mathbf{1}^{st}$	$\mathbf{18}^{th}$	$\mathbf{8}^{th}$	$\mathbf{66}^{th}$	$\mathbf{29}^{th}$	$\mathbf{8}^{th}$
...	...	...	...	...	...	...
Portal 80	$\mathbf{104}^{th}$	$\mathbf{34}^{th}$	$\mathbf{60}^{th}$	$\mathbf{4}^{th}$	$\mathbf{42}^{st}$	$\mathbf{52}^{st}$
...	...	...	...	...	...	...

A prior alternative has a longer distance to NIS and a shorter distance to PIS. Consequently, the closeness coefficient to the ideal solution for each alternative can be formulated as in Eq. 13, where $R(\mathrm{A}_l)$ denotes the final performance score of open portal l. The larger the $R(\mathrm{A}_l)$ score, the better the meta data quality published on portal l. The overall ranking of the monitored portals can therefore be generated based on the set of $R(\mathrm{A}_l)$ performance scores. Nonetheless, let us note that in Eq. 11 and 12, if:

- $\mathrm{Q}_{x(h)} = \{\mathrm{Q}_{u(\mathrm{core,res,extra})}, \mathrm{Q}_{c(\mathrm{core,res,extra})}, \mathrm{Q}_{o(\mathrm{F,L})} \dots$: a single and overall ranking of the portals is generated, i.e. all dimensions are aggregated into a unique and final score (see "Overall Ranking" in Table 3);
- $\mathrm{Q}_{x(h)} = \{\mathrm{Q}_{u(\mathrm{core,res,extra})}\}$ or $\{\mathrm{Q}_{c(\mathrm{core,res,extra})}\}$ or $\dots$: one ranking per quality dimension (i.e., Q_u, Q_c, Q_o, Q_a and/or Q_r) is generated (see "One ranking per quality dimension" in Table 3).

4. USE CASE

The objective of this use case is to present how, in practice, the ODPQ dashboard and associated widgets can benefit end-users such as governments, municipalities, or still developers for creating innovative services and benchmarks on top of it. Figure 4 presents the overall architecture, giving insight into how "Backend system" and "Web/User Interfaces" (databases, portals, end-users...) interact with each other. The architecture differentiates the "Open Data Portal Watch" components developed in our previous work [19] – *which crawls and independently assesses CKAN Open Data portal quality (see ① to ④ in Figure 4)* – with the OPDQ dashboard, which rather tackles the MCDM comparison problem (see ⑤ to ⑨). When an end-user requests for the Open Data portal quality comparison service (see ⑤), the ODPQ backend system retrieves – *through a RESTful API; see ⑥* – the data quality metrics computed by the "Open Data Portal Watch" and starts the AHP-based comparison process (see ⑦). It is worth noting that end-users can also subscribe to the service by specifying the subscription interval and duration[8] (e.g., if the end-user wants to receive the portal ranking on a daily, weekly or monthly basis). Finally, since OPDQ comparisons are carried out at different intervals of time, it is also possible to compute the ranking and quality evolution of the portals over time, as emphasized with ⑧ in Figure 4.

The following sections give insight into stages ⑤ to ⑧ where, as of January 2016, over 900K datasets have been reg-

[8]To support such a functionality, recent IoT standards have been implemented, namely the O-MI and O-DF standards [8].

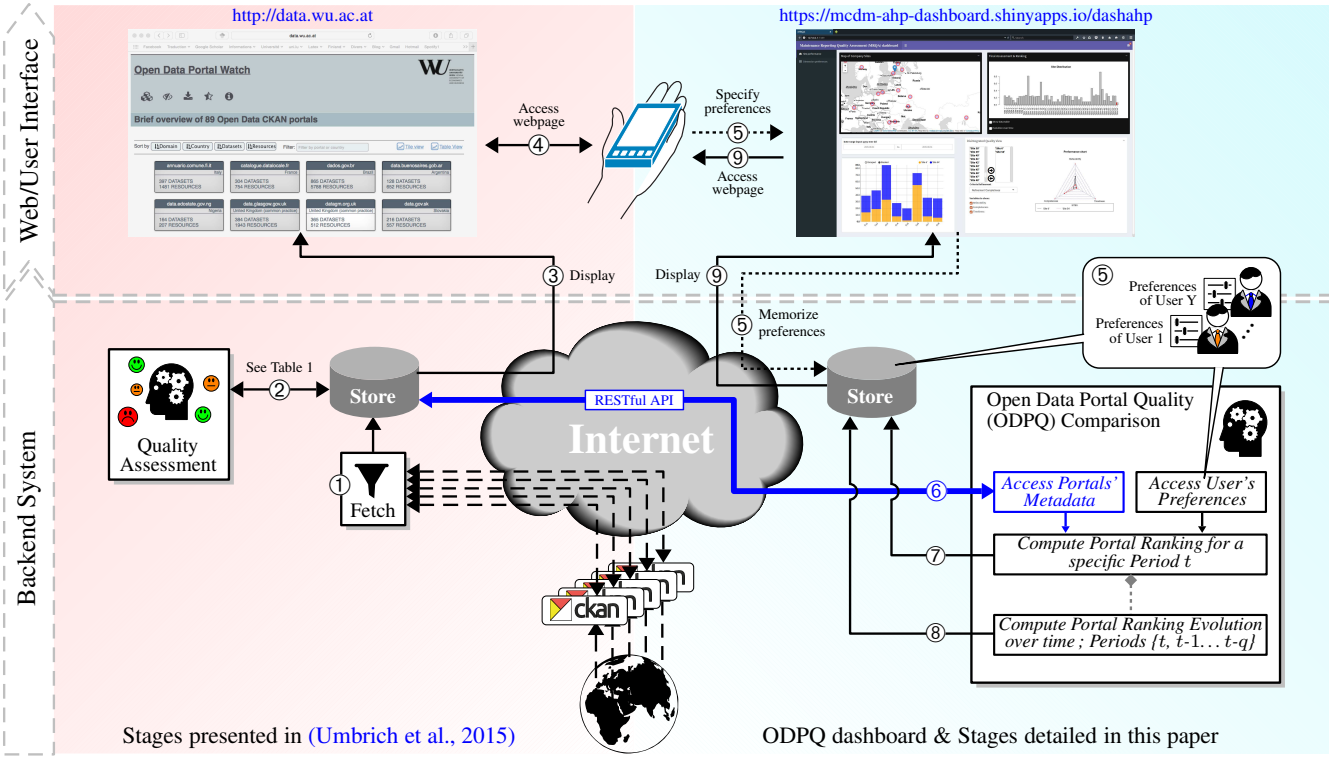


Figure 4: Overall infrastructure underlying the OPDQ dashboard

Table 4: Matching of Portal Names/Numbers

Portal N^o	URL	Rank	Country
70	daten_rlp_de	36^{th}	Deutschland
88	govdata_de	29^{th}	Deutschland
92	linked. . . uni-mannheim.de	112^{th}	Deutschland
93	open_nrw	143^{rd}	Deutschland
95	offenedaten_de	70^{th}	Deutschland
102	opendata_bayern_de	12^{th}	Deutschland
123	transparenz_hamburg_de	145^{th}	Deutschland
131	www_daten_rlp_de	37^{th}	Deutschland
137	www_offene	70^{th}	Deutschland
138	www_opendata-hro_de	10^{th}	Deutschland
144	www_opengov-muenchen_de	18^{th}	Deutschland
42	data_lexingtonky_gov	3^{rd}	UK
52	data_ottawa_ca	2^{nd}	Canada
53	data_overheid_nl	1^{st}	Netherlands

istered across the 146 CKAN portals monitored in our system. In total, those portals are spread across 44 countries, whose greatest majority are located in US (14%), Germany and UK (8%) and Spain (6%). These portals are referred to as *portal 1,..,146* in this paper, but the reader can refer to the URL provided in Figure 6 to identify what CKAN portal corresponds to what number (Open Data portals discussed in our results have nonetheless been listed in Table 4). Section 4.1 discusses the Open Data comparison results for a specific period and considering a specific set of user preferences. Section 4.2 gives further insight into the evolution (over time) of the quality of the monitored portals.

4.1 Equivalence between Quality Dimensions

The end-user wants to compare the quality of the CKAN Open Data portals without prioritizing any quality dimension. To this end, the user performs pairwise comparisons by specifying that all quality dimensions (including sub-dimensions) are equal in importance. Figure 5 provides insight into the different widgets/UIs and associated functionalities supported by the OPDQ dashboard[9], namely: *(i) Preference specification view:* provides end-users with the possibility to modify their preferences with regard to the quality dimension importance; *(ii) Histogram view:* displays the final quality score and ranking of the monitored portals at a specific point in time (e.g., a specific week); *(iii) Polar Chart view:* enables the selection of a set of Open Data portals in order to analyze/visualize how those portals behave with respect to one or more quality dimensions; and *(iv) Map view:* shows the Open Data portal locations.

Figure 6 shows the *Histogram view* related to our scenario, where the x-axis refers to the 146 portals and the y-axis to the quality score obtained after applying AHP. The first observation is that three Open Data portals stand out with a quality score ≥ 0.25 (see portals 53, 42 and 52 respectively), while the other portals have lower quality scores (most of them lying between 0.10 and 0.25)[10]. In this scenario, let us consider that the end-user is particularly interested in investigating the quality of German Open Data portals (e.g., for governmental survey purposes). As a first observation, the histogram emphasizes (with red/dashed shapes in Figure 6) that all German portals have quality scores varying from around 0.12 to 0.28 , and a final ranking between 18 and 145 (not directly visible on Figure 6 but highlighted in

[9]The OPDQ dashboard is available at the following URL: https://mcdm-ahp-dashboard.shinyapps.io/dashahp/

[10]A quality score of 0 means that the portal was unreachable at the time the portal was crawled

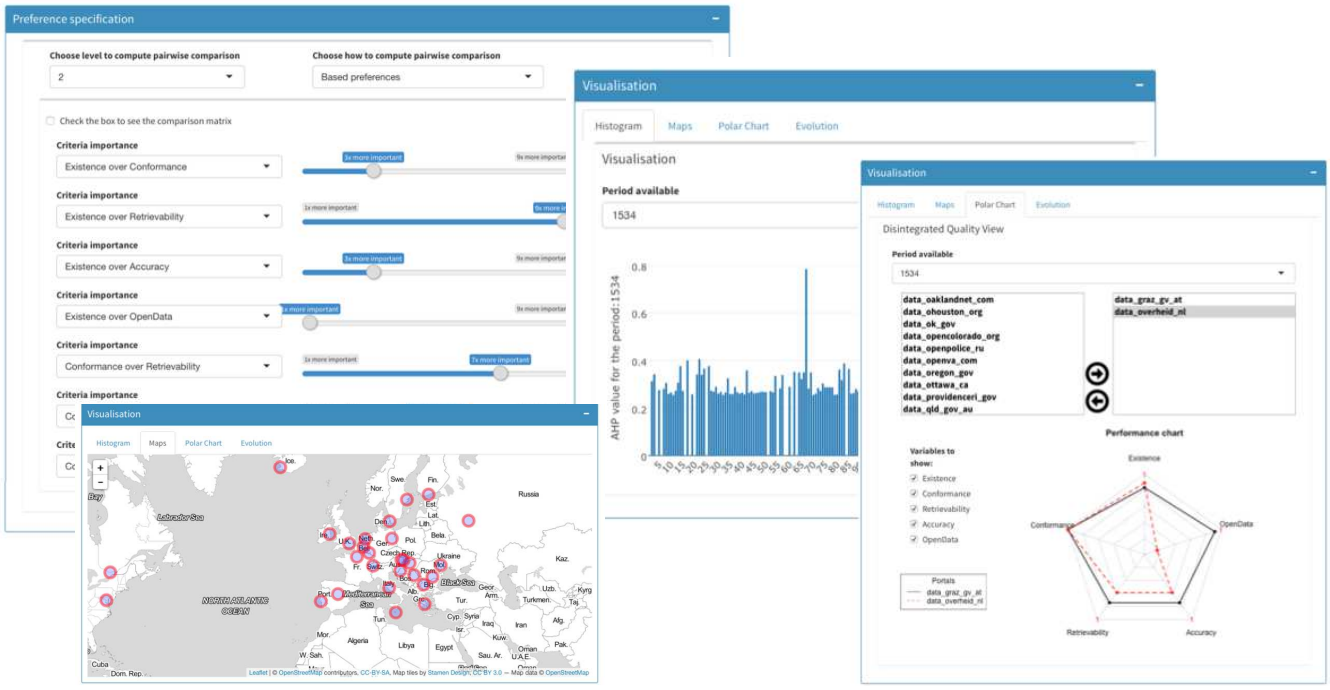
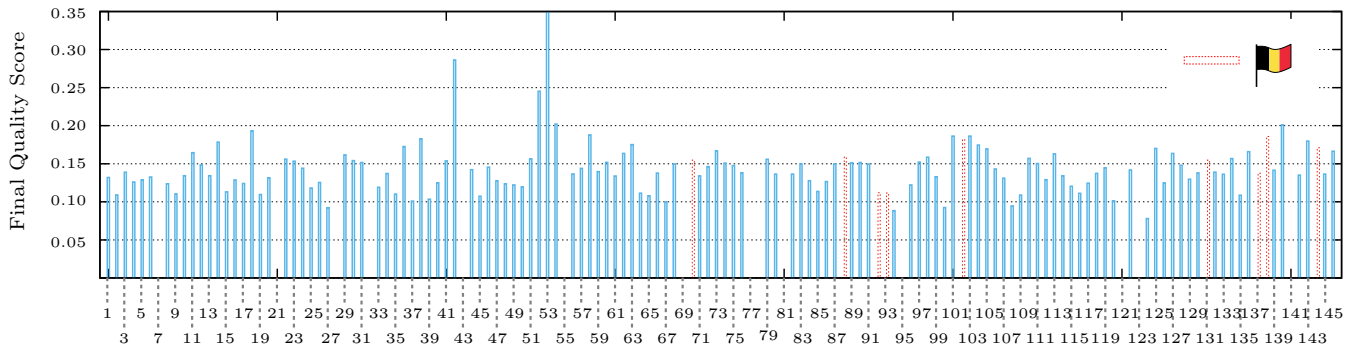


Figure 5: Screenshots of the ODPQ dashboard, including the comparison study of the german portals



Portal Number – portals' name available at: http://mcdm-ahp-dashboard.shinyapps.io/dashahp_CKAN

Figure 6: *Histogram view:* Open Data Portal Quality score obtained after applying AHP (week 53, 2015)

Table 4). The end-user now uses the *Polar Chart view* (see Figure 7) to compare three of those German portals, which correspond to city open data portals, namely portals 92, 138, 144. The Polar Chart highlights that one out of the three city portals (namely portal 138) is well ranked with respect to four of the five quality dimensions defined at level 2 of the AHP structure. Portal 144 is in the middle range of the ranking regarding the Retrievability dimension and in lower part of the ranking regarding the Openness dimension. Considering the last city portal (portal 92), it distinguishes itself from the two other portals with the Retrievability dimension (ranked in the top 10), although it is not high-ranking regarding the four other dimensions.

The end-user could potentially refine a step further those observations by, in a similar manner, 'disaggregating' each level 2's quality dimension to understand how specific portals behave with respect to sub-dimensions (i.e., level 3's dimensions). However, this is not discussed in this paper.

4.2 Evolution of portal quality and ranking

The end-user is now interested in studying how the German portals' quality evolves over time. To this end, a widget has been developed (see Figure 8) to display the portal ranking (x-axis) according to the number of datasets (y-axis) as well as the resources available on each portal (bubble size).

The portal evolution computation still considers the previous set of dimension preferences, i.e. all quality dimensions (including sub-dimensions) are equal in importance. Figures 8(a) and 8(c) gives insight into the quality comparison results over the experimental/monitoring period (i.e., over weeks 44-53 of the year 2015). First, it appears here that the number of datasets and resources does not directly impact on the final ranking, e.g. portal 88 has the highest number of datasets/resources among the German open data portals and is well ranked over this period of time (even ranked first at week 53). Second, it can be noted that the ranking among the German portals evolves over time,

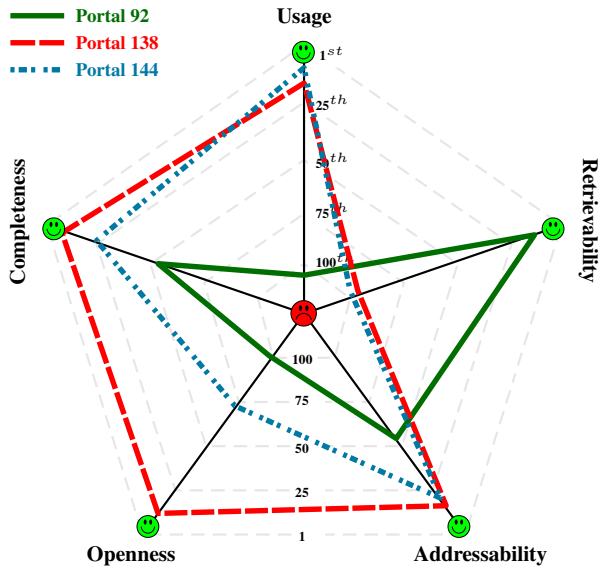


Figure 7: German Portal Comparison (Week 44)

and the reason for that is twofold: *i)* one or several portals published (or lost) information/datasets/resources over the period of time, which results in moving up or down some of the German portals in the ranking (e.g., Figures 8(a) shows that portal 88 lost hundreds of datasets compared with week 44/Figures 8(a)); *(ii)* one or several portals were momentarily unreachable, which results in moving down significantly the unreachable portal in the ranking.

Let us now consider a second scenario where the end-user attaches greater importance to the Openness dimension (Q_o), and particularly regarding Licences (i.e., $Q_o(L)$). The end-user thus specifies – *using the Preference specification view (see Figure 5)* – that $Q_o(L)$ is strongly more important (9 on Saaty's scale) than $Q_o(F)$. Figures 8(b) and 8(d) show the ranking evolution of the German portals the same two periods of time. It can be noted here that the ranking is much different from the previous scenario; for example, portal 88 that was (in the previous scenario) ranked 3^{rd} and 1^{st} respectively for weeks 44 and 53, is now (in this second scenario) ranked 8^{th}. Another observation that can be made between scenarios 1 and 2 is that, while rankings vary substantially between weeks 44 and 53 in scenario 1, they almost remain unchanged in scenario 2, which means that the openness dimension did not play an important role in the ranking evolution observed in scenario 1. In summary, all these observations show how carefully the results must be interpreted according to the set of preferences specified by the user.

5. CONCLUSION & DISCUSSION

Organizations from all over the world are under increasing pressure to release, in an open and transparent manner, their data to a variety of users (citizens, businesses, academics...). Data openness is expected to improve the decision making of both governments and individuals. Although opportunities are wide and worth exploring, first critical voices and reports claim that there is a quality problem in Open Data, which has a non-negligible impact on open and *e*-government initiatives. Our study points out the lack of frameworks and tools to dynamically assess Open Data portal quality and compare those portals with one another.

To address this lack, along with the multi-criteria decision making (MCDM) nature of the comparison process, our research applies the Analytic Hierarchy Process (AHP) technique, whose methodology is turned into an Open Data Portal Quality (ODPQ) Web dashboard that enables any Open Data stakeholder to identify, at any point in time, the quality and ranking of one or a group of Open Data portals. A use case, in which 146 CKAN portals (and over $900K$ datasets) are monitored, is presented showing how end-user preferences can be taken into consideration in the AHP-based comparison process. To put it another way, this use case shows how open data end-users can benefit from various widgets supported by the OPDQ dashboard (see e.g. Figure 5), thus opening up opportunities to build on top of it innovative *e*-government services and benchmarks.

In terms of research perspectives, the AHP structure will likely be extended by including new quality dimension based on a thorough literature review, e.g. to take into consideration the quantity of datasets/resources on an Open Data portal (the more datasets, the more datasets might not have the same). Another perspective of this work is to propose a generalized metadata schema, which would be able to map metadata schemas observed on CKAN, Socrata and other Open Data software frameworks to metadata standards such as W3C's DCAT (Data Catalog Vocabulary)[11]. This mapping is intended as a homogenization of different metadata sources by using the DCAT vocabulary. Finally, dealing with uncertainty is also an important aspect to be considered in future research work when computing the quality metrics (e.g., by combining Fuzzy logic with AHP) [11].

6. ACKNOWLEDGMENTS

The research leading to this publication is supported by the EU's H2020 Programme (grant 688203), the National Research Fund Luxembourg (grant 9095399), as well as the Austrian Research Promotion Agency (grant 849982).

7. REFERENCES

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

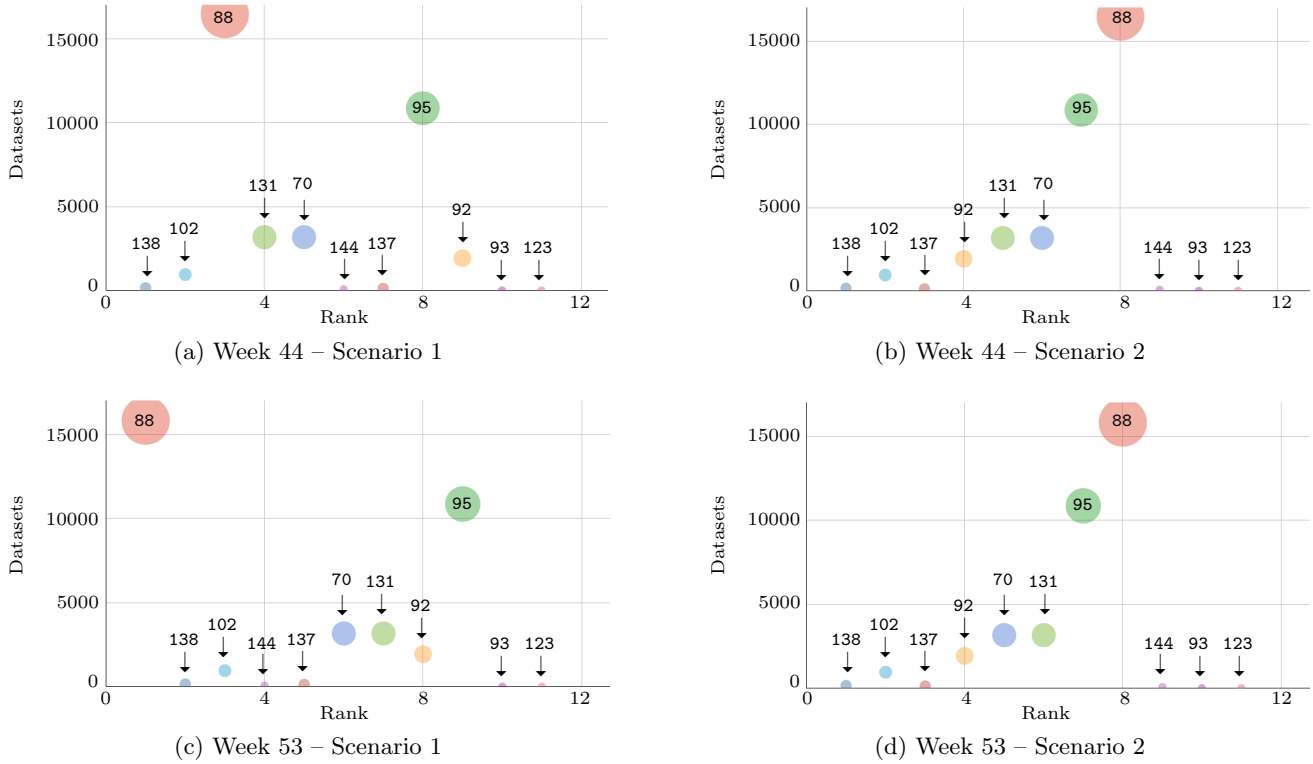
[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):1–52, 2009.

[3] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic. A metrics-driven approach for quality assessment of linked open data. *Journal of theoretical and applied electronic commerce research*, 9(2):64–79, 2014.

[4] A. L. Blumenthal. *The process of cognition*. Prentice Hall/Pearson Education, 1977.

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

[11] DCAT is a W3C metadata recommendation for publishing data on the Web: http://www.w3.org/TR/vocab-dcat/



(a) Week 44 – Scenario 1

(b) Week 44 – Scenario 2

(c) Week 53 – Scenario 1

(d) Week 53 – Scenario 2

Figure 8: Evolution of ranking vs datasets for the week 44 and 53 of the year 2015

[6] P. Conradie and S. Choenni. On the barriers for local government releasing open data. *Government Information Quarterly*, 31:S10–S17, 2014.

[7] J. Figueira, S. Greco, and M. Ehrgott. *Multiple criteria decision analysis: state of the art surveys*. Springer Science & Business Media, 2005.

[8] K. Främling, S. Kubler, and A. Buda. Universal messaging standards for the iot from a lifecycle management perspective. *IEEE Internet of Things Journal*, 1(4):319–327, 2014.

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

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

[11] S. Kubler, A. Voisin, W. Derigent, A. Thomas, E. Rondeau, and K. Främling. Group fuzzy ahp approach to embed relevant data on "communicating material". *Computers in Industry*, 65(4):675–692, 2014.

[12] J. Kučera, D. Chlapek, and M. Nečaskỳ. Open government data catalogs: Current approaches and quality perspective. In Springer, editor, *Technology-Enabled Innovation for Democracy, Government and Governance*, pages 152–166, 2013.

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

[14] A. Molnar, M. Janssen, and V. Weerakkody. e-government theories and challenges: findings from a plenary expert panel. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, 2015.

[15] J. L. Mumpower, L. D. Phillips, O. Renn, and V. R. R. Uppuluri. *Expert Judgment and Expert Systems*, volume 35. Springer Science & Business Media, 2012.

[16] K. J. Reiche, E. Höfig, and I. Schieferdecker. Assessment and visualization of metadata quality for open government data. In *Conference for E-Democracy and Open Governement*, 2014.

[17] T. L. Saaty. *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

[18] T. L. Saaty. *Decision making with dependence and feedback: The analytic network process*, volume 4922. RWS publications Pittsburgh, 1996.

[19] J. Umbrich, S. Neumaier, and A. Polleres. Quality assessment & evolution of open data portals. In *3rd International Conference on Future Internet of Things and Cloud*, pages 404–411, Roma, Italy, 2015.

[20] A. Zuiderwijk and M. Janssen. The negative effects of open government data - investigating the dark side of open data. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, pages 147–152, 2014.

[21] A. Zuiderwijk, M. Janssen, K. Poulis, and G. van de Kaa. Open data for competitive advantage: insights from open data use by companies. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, 2015.