Characterizing Open Government Data Available on the Web from the Quality Perspective: A Systematic Mapping Study

Rafael Chiaradia Almeida[a], Glauco de Figueiredo Carneiro[b] and Edward David Moreno[c]
Programa de Pós-Graduação em Ciência da Computação (PROCC), Federal University of Sergipe (UFS), Sao Cristovao, Sergipe, Brazil

Keywords: Open Government Data, Systematic Mapping, Data Quality Evaluation.

Abstract: Context: Data openness can create opportunities for new and disruptive digital services on the web that has the potential to benefit the whole society. However, the quality of those data is a crucial factor for the success of any endeavor based on information made available by the government. Objective: Analyze the current state of the art of quality evaluation of open government data available on the web from the perspective of discoverability, accessibility, and usability. Methods: We performed a systematic mapping review of the published peer-reviewed literature from 2011 to 2021 to gather evidence on how practitioners and researchers evaluate the quality of open government data. Results: Out of 792 records, we selected 21 articles from the literature. Findings suggest no consensus regarding the quality evaluation of open government data. Most studies did not mention the dataset´s application domain, and the preferred data analysis approach mainly relies on human observation. Of the non-conformities cited, data discoverability and usability outstand from the others. Conclusions: There is also no consensus regarding the dimensions to be included in the evaluation. None of the selected articles reported the use of machine learning algorithms for this end.

1 INTRODUCTION

With the evolution of digital technology, a large amount of data has been publicly available (Utamachant and Anutariya, 2018). Among those data is the open data, as the name reveals, is open to the public (Yi, 2019) and can be freely used, modified, and shared by anyone for any purpose (Vetrò et al., 2016). According to *Open Data Barometer*, open data plays a relevant role in ensuring effective institutions and securing public access to government information (Brandusescu et al., 2018). Open government data (OGD) refers to data provided to the public by government institutions (Kassen, 2013). Despite the significant proliferation of platforms to make these data available, quality issues are a crucial factor (Kučera et al., 2013a) for the success of OGD initiatives (Belhiah and Bounabat, 2017). For example, the discrepancy in the terms and data types negatively affects reuse and its effective use (Zuiderwijk et al., 2016).

Discovering the relevant data is a prerequisite (Kučera et al., 2013a) for a user to be able to use open data. When a dataset is publicly online, users must access it without barriers (Dander, 2014), either through a file download, a portal query, or API access, among other possibilities. Moreover, datasets with non-conformities like missing values, outdated information, and inappropriate metadata, may not fit the user needs (Máchová and Lněnička, 2017). For those reasons, discoverability, accessibility, and usability are essential dimensions for the quality evaluation of open government data.

This Systematic Mapping is part of a larger joint project whose goal is to propose a roadmap for open government data quality evaluation. As the first step of this project, we endeavor to analyze the current state of the art in open government data quality evaluation. The Research Question (RQ) of this Systematic Mapping is as follows: ”What techniques, frameworks, and machine learning algorithms have researchers and stakeholders used to evaluate the quality of open government data regarding its discoverability, accessibility, and usability”? This research question is in line with the goal of this review. Researchers have conducted several studies to evaluate the quality of OGD (Oliveira et al., 2016) but, according to Sadiq and Indulska (2017), there is still a gap

[a] https://orcid.org/0000-0001-8349-8294
[b] https://orcid.org/0000-0001-6241-1612
[c] https://orcid.org/0000-0002-4786-9243

Almeida, R., Carneiro, G. and Moreno, E.
Characterizing Open Government Data Available on the Web from the Quality Perspective: A Systematic Mapping Study.
DOI: 10.5220/0011585300003318
In *Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022)*, pages 214-225
ISBN: 978-989-758-613-2; ISSN: 2184-3252

in the evaluation of data quality dimensions.

To answer the proposed research question a Systematic Mapping Study (SMS) has been carried out to gather evidence provided by papers published in peer-reviewed conferences and journals from 2011 to 2021. We initially found 792 papers as a result of the applied search strings in specific electronic databases and the execution of snowballing procedure (Wohlin, 2014), from which we considered 21 studies as relevant. Findings suggest that there is no standardization in data quality evaluation, but some dimensions have been more used in studies. We also found that there are dimensions with different names but with related quality evaluation purposes, and that's the reason we decided to create dimension groups. Those groups can have different influences on the three main dimensions: discoverability, accessibility, and usability. Regarding the data sources used in open government data quality evaluation studies, most did not mention the application domain of datasets used in the analysis. The approach used by most of the articles in data analysis was human observation. And most of the non-conformities cited by the selected articles were related to data discoverability and usability.

The remainder of this paper is organized as follows. Section 2 presents the design we adopted to conduct this systematic mapping study. Section 3 presents the key findings to the stated research question and Section 4 concludes this work.

2 RESEARCH DESIGN

We performed a systematic mapping review of the published peer-reviewed literature, from 2011 to 2021, to gather existing evidence of the techniques, frameworks, or machine learning algorithms that have been used to evaluate the quality of open government data (OGD). Systematic mapping is a type of secondary study that has the goal to describe the extent of the research in a field and identify gaps in the research base. It identifies gaps in the research where further primary research is needed, and areas where no systematic reviews have been conducted and there is scope for future review work (Clapton et al., 2009).

Systematic mapping provides descriptive information about the state of the art of a topic and a summary of the research conducted in a specific period (Clapton et al., 2009). The overall process for the selection of relevant studies is presented in Table 1 and described in more detail in the following subsections.

Table 1: Steps for the selection process.

Step	Description
(1)	Apply the search strings to obtain a list of candidade papers in specific electronic databases
(2)	Remove replicated papers from the list
(3)	Apply the exclusion criteria in the listed papers
(4)	Apply the inclusion criteria after reading abstracts, introduction and conclusion in papers not excluded in step 3
(5)	Apply quality criteria in selected papers from step 4

2.1 Planning

We conducted this SMS based on a protocol comprised of objectives, research questions, selected electronic databases, search strings, and selection procedures comprised of exclusion, inclusion, and quality criteria to select studies from which we aim to answer the stated research questions (Wohlin et al., 2012). The goal of this study is presented in Table 2 according to the Goal Question Metric (GQM) approach (Basili and Rombach, 1988).

Table 2: The goal of this SMS according to the GQM approach.

Analyze	Open Government Data
for the purpose of	quality evaluation
with respect to	usability, accessibility and discoverability
from the point of view of	researchers and stakeholders
in the context of	techniques, frameworks and machine learning algorithms

The Research Question (RQ) is *"What techniques, frameworks, and machine learning algorithms have researchers and stakeholders used to evaluate the quality of open government data regarding its discoverability, accessibility, and usability"?* The motivation behind RQ is justified by the fact that high-quality data can be used by many institutions (public or private) to improve business processes, make smart decisions and create strategic advantages (Behkamal et al., 2014). On the other hand, a government that publishes data with low quality, such as missing metadata or duplicated fields, can create a bad reputation among its citizens (Kubler et al., 2018a).

The specific research questions have the goal to gather evidence to support the answer to the stated RQ. This research question is in line with the goal of this review and has been derived into four specific research questions, as follows. Specific Research Question 1 (SRQ1): *What are the key quality dimensions adopted by researchers to evaluate the quality of open government data?* Data quality is a multidimensional construct and can be analyzed under different perspectives, it is important to understand how researchers are taking into consideration those different angles of analysis. Specific Research Question 2

(SRQ2): *What kind of data sources are used in open government data quality evaluation studies?* The data sources are a crucial component of any research project from which information can be extracted to unveil trends and patterns that otherwise would not be known by the research community. Specific Research Question 3 (SRQ3): *What is the approach used in data analysis?* Understanding how the data are being evaluated in practice is critical and will enrich the research. Specific Research Question 4 (SRQ4): *What are the data quality non-conformities found by researchers?* It is essential to know the existing problems so that interested parties can take steps to resolve them.

We considered the PICO criteria (Stone, 2002) to define the search string, as shown in Table 3. The search strings are based on this criteria for the selective process of papers for this review. The steps to build a search string to identify studies in the target repositories are shown in Table 4 and 5. The Table 4 refers to major terms for the research objectives. We also considered the use of alternative terms and synonyms for these major terms. For example, the term machine learning can be associated with terms such as artificial intelligence and deep learning. These alternative terms, as shown in Table 5, can be also included in the search string. We built the final search string by joining the major terms with the Boolean ”AND” and joining the alternative terms to the main terms with the Boolean ”OR”. The focus of the formed search string is to identify studies targeting the research question of this systematic mapping.

Table 3: PICO criteria for search strings.

(P)opulation	papers focusing on open government data analysis
(I)ntevention	evaluation methods used to analyze open government data usability
(C)omparison	differences between methods using machine learning and other frameworks
(O)utcomes	usability status of open data portals and best usability evaluation methods

Table 4: Major terms for the research objectives.

Criteria	Major Terms
(P)opulation	AND "open government data"
(I)ntevention	AND "usability evaluation"
(C)omparison	AND "machine learning"
(O)utcomes	AND "methods"

Table 6 presents the electronic databases from which we retrieved the papers along with the respective search strings used in that process. The target databases were ACM Digital Library, IEEE Digital Library, and Scopus. All searches were performed on October 10, 2021.

Table 5: Alternative terms from major terms.

Major Terms	Alternative Terms
AND "open government data"	("OGD" OR "open data portals")
AND "usability evaluation"	("quality evaluation" OR "data evaluation" OR "data quality" OR "reusability" OR "accessibility" OR "discoverability")
AND "machine learning"	("artificial intelligence" OR "AI" OR "deep learning" OR "data management" OR "data science" OR "data analysis" OR "data engineering")
AND "methods"	("frameworks" OR "techniques" OR "algorithms")

Table 6: Electronic databases selected for this Systematic Mapping.

Database and URL	Search Strings
ACM Digital Library portal.acm.org	("open government data" OR "OGD" OR "open data portals") AND ("usability evaluation" OR "quality evaluation" OR "data evaluation" OR "data quality" OR "reusability" OR "accessibility" OR "discoverability") AND ("machine learning" OR "artificial
IEEE Digital Library ieeexplore.ieee.org	("open government data" OR "OGD" OR "open data portals") AND ("usability evaluation" OR "quality evaluation" OR "data evaluation" OR "data quality" OR "reusability" OR "accessibility" OR "discoverability") AND ("machine learning" OR "artificial intelligence" OR "AI" OR "deep learning" OR "data management" OR "data science" OR "data analysis" OR "data engineering") AND ("methods" OR "frameworks" OR "techniques" OR "algorithms")
Scopus www.scopus.com	("open government data" OR "OGD" OR "open data portals") AND ("usability evaluation" OR "quality evaluation" OR "data evaluation" OR "data quality" OR "reusability" OR "accessibility" OR "discoverability") AND ("machine learning" OR "artificial

Table 7 presents the criteria for exclusion, inclusion, and quality evaluation of papers in this review. The OR connective used in the exclusion criteria means that the exclusion criteria are independent, i.e., meeting only one criterion is enough to exclude the paper. On the other hand, the AND connective in the inclusion criteria means that all inclusion criteria must be met to select the paper under analysis. Table 7 also presents the quality criteria used for this review represented as questions adjusted from their original version from Dyba and Dingsoyr (Dybå and Dingsøyr, 2008). We evaluated all the remaining papers that passed the exclusion and inclusion criteria using the quality criteria presented in the same table. All these criteria must be met (i.e., the answer must be YES for each one) to permanently select the paper, otherwise, the paper must be excluded. The exclusion, inclusion, and quality criteria were used in the steps for the selection process as already presented in Table 1. According to Table 8, at the end of the selection process, all the retrieved papers were classified in one of the three options: *Excluded, Not Selected* and *Selected*.

2.2 Execution

The quantitative evolution of the selection process execution is summarized in Figure 1.

The figure uses the PRISMA flow diagram (Moher et al., 2009) and shows the performed steps and the respective number of papers for each phase of the

Table 7: Exclusion, inclusion and quality criteria.

Type	Id	Description	Connective or Answer
Exclusion	E1	Published earlier than 2011	OR
Exclusion	E2	The paper does not present a primary study	OR
Exclusion	E3	The paper has less than 4 pages	OR
Exclusion	E4	The paper is not written in english	OR
Exclusion	E5	The paper was not published in a peer-reviewed journal or	OR
Inclusion	I1	The study should be conducted within the scope of open data portals	AND
Inclusion	I2	The paper should present a study on open government data analysis	AND
Quality	Q1	Are the aims of the study clearly specified?	YES/NO
Quality	Q2	Is the context of the study clearly stated?	YES/NO
Quality	Q3	Does the research design support the aims of the study?	YES/NO
Quality	Q4	Does the study have an adequate description of the methods used to analyze open government data?	YES/NO
Quality	Q5	Is the data analysis of the study rigorous and based on evidence or theoretical of reasoning instead of non-justified or ad hoc statements?	YES/NO

Table 8: Classification options for each retrieved paper.

Classification	Description
Excluded	Papers met the exclusion criteria.
Not Selected	Papers not excluded due to the exclusion criteria, but did not met the inclusion or quality criteria.
Selected	Papers did not meet the exclusion criteria and met both the inclusion and quality criteria.

systematic mapping.

According to Table 1, as a result of the execution of Step 1 (execution of the search string), we retrieved from the three selected repositories a total of 792 papers (Identification Phase of Figure 1). Considering that 55 papers were duplicated, we evaluated 737 regarding the alignment of their titles and abstracts to the stated specific research questions (Screening Phase of Figure 1).

The result of this evaluation was the exclusion of 652 papers and the inclusion of 85 papers, following the exclusion and inclusion criteria respectively already presented in Table 7. In the Eligibility Phase of Figure 1, we evaluated 85 papers to decide that 64 papers should not have been selected due to not meeting the quality criterion presented in Table 7. The final set of studies to answer the specific research questions is comprised of 21 papers (Included Phase of Figure 1).

Table 9 presents the effectiveness of the search strings considering the 792 retrieved papers. The repository database that most contributed to selected

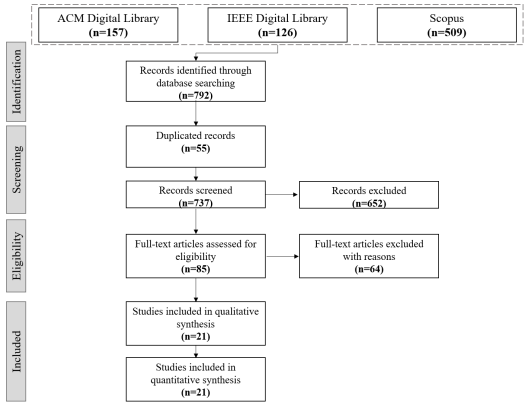

Figure 1: Phases of the selection process in numbers (adjusted from Moher et al. (2009)).

studies was the *Scopus* with 12 papers, corresponding to search effectiveness of 2,36%. The 21 selected studies represented 2,65% of all 792 papers retrieved by the search string.

Table 9: Effectiveness of the search strings.

Database	Papers Retrieved by the Search String	Selected Papers	Search Effectiveness
ACM Digital Library	157	7	**4,46%**
IEEE Digital Library	126	2	**1,59%**
Scopus	509	12	**2,36%**
TOTAL	792	21	**2,65%**

3 RESULTS

All the relevant information that was extracted from the 21 selected articles to answer the four specific research questions (SRQ1, SRQ2, SRQ3, and SRQ4) are presented in a form of a complete table that can be downloaded in an Excell format through Zenodo (https://doi.org/10.5281/zenodo.7015916). Figure 2 summarizes the evidence where each branch has the associated SRQ. The total amount of references for each of the four branches of Figure 2 does not correspond to the total of 21 analyzed articles. The reason for this difference is that most of the studies fall into more than one of the available groups. The branch named Quality Dimensions Analyzed (SRQ1) indicates not only the quality dimensions that were presented in each article but also if that dimension is used to analyze data and/or metadata which is represented by a letter D (data) and/or M (metadata) beside each dimension name. Also, in that same branch, the dimension that has different names but has the same quality evaluation objective were placed together in a branch.

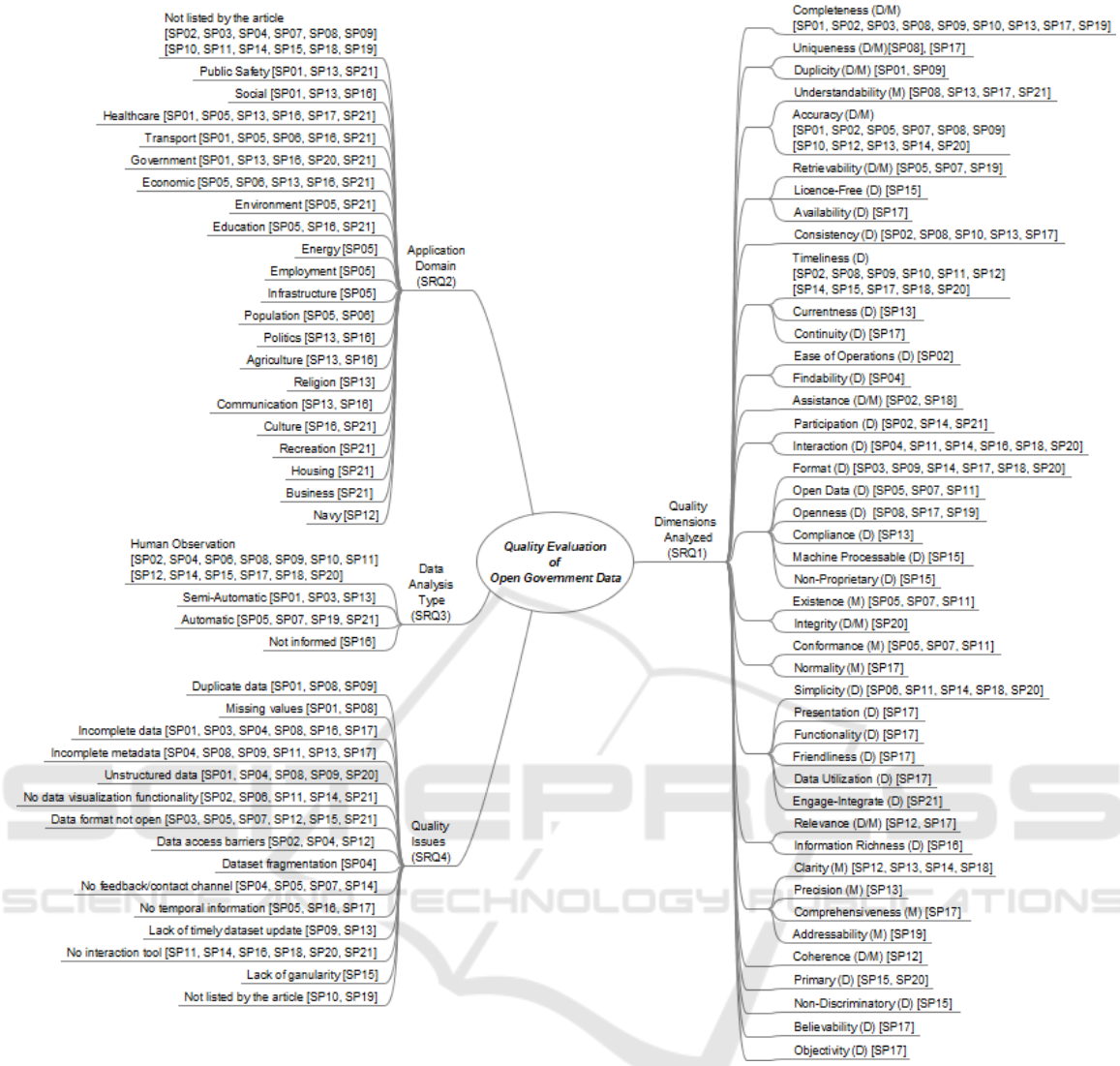

Figure 2: Evidence from the literature to answer specific research questions.

In this section we will discuss in detail those gathered information and present our findings. With the purpose to make it in a clearer and more comprehensive way, we divided this section into four, each one representing a specific research question.

3.1 SRQ1 - What Are the Key Quality Dimensions Adopted by Researches to Evaluate the Quality of Open Government Data?

During the SMS we noticed that different dimensions were being used no analyse datasets and, sometimes, with different nomenclature but the same quality evaluation objective. That finding was also reported by some of the selected articles. Sadiq and Indulska (2017), for example, quoted that one of the challenges in open data quality management is a common understanding of the data quality dimension. According to Stróżyna et al. (2018) there is no consensus among academics regarding an approach to the assessment of data quality.

This lack of pattern impairs the data quality analysis, but, on the other hand, it is difficult to create a standard that embrace every scenario. Neumaier et al. (2016) states that the selection of a proper set of quality dimensions to evaluate datasets is highly context specific since their purposing is testing the fitness for use of data for a specific task; and that's the reason why quality dimensions differ among the data quality methodologies (Batini et al., 2009).

Figure 2 plots in its right side data related to the Quality Dimensions Analyzed (SRQ1). We identified 49 different names for the quality dimensions, but 5 of them were excluded because they are broader terms, that is, they can not be measure without analyzing other dimensions. These are: Accessibility, Usability, Discoverability, Structuredness and Trust. Of the remainder dimensions, 24 has the same data quality evaluation objective of others and for that reason they were placed together in a branch. Eight dimensions branch were cited only 3 times or less and for that reason they were not considered relevant. These branches and their frequency of appearance are: Relevance and Information Richness (3 studies - 14%), Primary (2 studies - 10%), Ease of Operations and Findability (2 studies - 10%), Assistance (2 studies - 10%), Coherence (1 study - 5%), Non-Discriminatory (1 study - 5%), Believability (1 study - 5%) and Objectivity (1 study - 5%). We ended up with 12 quality dimensions group, each one containing one or more related dimensions. Table 10 presents those groups in order of most appearance in the articles.

It is important to report that 5 articles (SP09, SP11, SP14, SP18, SP20) analyzed characteristics of the data (or metadata) without mentioning the name of the dimension. In that case we had to convert them into dimensions according to the aspects that were evaluated. Also, 5 articles (SP02, SP06, SP11, SP20, SP21) used dimensions and metrics to evaluate general characteristics of the open data portal, but those were not considered because our work is about data quality evaluation and not portal quality evaluation.

In terms of number of dimensions evaluated, all the articles together analysed 127 dimensions, an average of 6 dimensions per article. Only one article evaluated more than 10 dimensions (SP17 - 21 dimensions) and only 3 articles analysed 3 or less dimensions (SP01 - 3 dimensions, SP03 - 2 dimensions, SP06 - 1 dimension).

When a dataset is being evaluated in terms of its quality, the data itself can be the target of the analysis, or the metadata, or both. In the selected studies, 27 dimensions used data as a primary source of analysis, 8 dimensions used metadata and 9 dimensions used both. This happens because for some dimensions the metadata, for example, will not give any quality information, or vice-versa. Take Timeliness as an example, the importance here is to analyze if the data is up to date. But for some dimensions both data and metadata will be an important source of information. In the case of Accuracy analysis, it is important to have a data that is precise and also a metadata that correctly describe the data that is being published.

Table 10: Dimensions groups found in the articles.

Dimension Group	Frequency
Accuracy Understandability	13 studies - 62%
Timeliness Currentness Continuity	12 studies - 57%
Format Open Data Openness Compliance Machine Processable Non-Proprietary	11 studies - 52%
Completeness	9 studies - 43%
Participation Interaction	8 studies - 38%
Simplicity Presentation Functionality Friendliness Data Utilization Engage-Integrate	7 studies - 33%
Clarity Precision Comprehensiveness Addressability	6 studies - 29%
Retrievability Licence-Free Availability	5 studies - 24%
Consistency	5 studies - 24%
Duplicity Uniqueness	4 studies - 19%
Existence Integrity	4 studies - 19%
Conformance Normality	4 studies - 19%

We will now present the definition of the dimensions presented in Table 10. Even though the dimensions Accessibility, Usability, Discoverability, Structuredness and Trust were not considered part of any quality dimension group, we decided to present the concept of those constructs and explain the reasons behind our decision of exclusion.

Accessibility was cited by 8 articles (SP02, SP04, SP12, SP13, SP15, SP17, SP20, SP21) and is defined by Stróżyna et al. (2018) as the the possibility to retrieve data from a source. When an user find a database, he should be able to access it without barriers. Other dimensions can affect the data accessibility, such as the format in which data are published, the search tool used, and the metadata of the dataset (Maali et al., 2010).

Usability appeared in 4 articles (SP04, SP16, SP17, SP19). According to Magalhães and Roseira (2016), this dimension can take many forms. Even though access to the data is not an issue, users can, for example, be partially or completely unable to explore the datasets available due to lack of necessary

resources or computational skills. That is the reason we decided to not consider it as part of any quality dimension group, because for a data (or metadata) to meet the usability criteria, it has first to attend others dimensions criteria, such as accessibility, format and completeness.

Discoverability appeared in only one article (SP16). At this research, Dahbi et al. (2018) highlights that an OGD is only really open if it can be easily discoverable, that is, users should be able to search and access relevant data in a simple and efficient way. The same way as Usability, to promote data Discoverability datasets must have other characteristics such as structured, complete and accurate metadata, which are other quality dimensions.

The last two broader dimensions that appeared on the selected articles are Structuredness (1 article - SP17) and Trust (1 article - SP21). According to Wu et al. (2021), Structuredness indicates the degree of organization of the dataset, and Zhu and Freeman (2019) developed a framework where the Trust dimension includes criteria such as completeness, currentness, availability, granularity and relevance.

The quality dimensions group presented on Table 10 were formed considering the quality evaluation objective of each dimension. When similarities or complementarities between those objectives were found, the dimensions were considered as related and, therefore, part of the same group. We will now conceptualize each one of the 12 groups and the definition of the dimensions itself will be sufficient to understand their relationship. As our research question is to analyse the dimensions used to evaluate the quality of open government data regarding its discoverability, accessibility and usability, it is fundamental to describe how the dimension groups affect those artefacts and that will be discussed along with their definition.

Accuracy (11 articles) is the extent to which a selected set of specific metadata keys accurately describe the resource (Sanabria et al., 2018), and its related dimension Understandability (4 articles) is attended by a dataset, according to Li et al. (2018), when the meaning of each column of data is clear. If you have correct and precise metadata within the data, it will positively affect the discoverability of the dataset. Not only by an ordinary user that make a search at a data portal or a search engine like Google, but also by an expert that will use an API or a SPARQL endpoint to find data. Usability will also be positively affected because the name and the meaning of each column, for example, is very important to understand and correctly make use of those data. Accessibility will not be affected.

Timeliness (11 articles) is achieved when data is made available to the public as soon as possible after the actual data is created, in order to preserve its value (Wang and Shepherd, 2020), and it has two related dimensions: Currentness (1 article), whose metrics of evaluation chosen by Utamachant and Anutariya (2018) were the timeliness after expiration (period between the publication of a dataset after the expiration of its previous versions) and the timeliness in a publication (period between the moment in which the dataset is available and its publication in the portal); and Continuity (1 article), that indicates whether there is continuity in the release of datasets on the same topic (Wu et al., 2021). When you have an updated dataset it will positively affect its usability. That's because citizens, organizations and researchers that are using those information in their projects depend on data that is made available in a continuous and timely manner, otherwise the information value is compromised. Discoverability can also be affected because if you are searching for an information of the current year, for example, and the dataset is two years out of date, you may have problems finding it. Accessibility will not be affected.

Format (6 articles) is a quality aspect that is attended when, according to Group (2007), the data available is machine-readable, provided in a convenient form, and offered without technological barriers for data consumers. Open Data (3 articles) dimension checks if the file format is based on an open standard, is machine readable and has an open license (Neumaier et al., 2016). Openness (3 articles) has the exactly the same definition of Open Data at Li et al. (2018). Compliance (1 article) appeared only at Neumaier et al. (2016) and it is a value (1-5) depending on dataset's file format according to Tim Berners-Lee's 5-star open data scheme. Machine Processable (1 article) and Non-Proprietary (1 article) appeared in the same work. The first is achieved when data is published in a structured manner to allow automated processing and the second when data is published in a format which is not controlled exclusively by a single entity. A data that is in a format that is not machine readable, pdf for an example, will not have associated metadata and that will make it difficult to be discovered. Using that same example, it is not an easy task to extract conclusions from data in a pdf format, that is, you will have problems to make simple statistics tasks such as calculating maximum, minimum or median of a dataset column, affecting both its accessibility and its usability.

Completeness (9 articles) is described by Kubler et al. (2016) as the extent to which the used data (or metadata) are non empty, that is, contain information,

as said by Sanabria et al. (2018). As in the Accuracy dimension, metadata is very important not only to explain the content of the dataset but also to make it easier to be found. So, the lack of metadata will affect dataset's discoverability and usability. Accessibility is not affected.

Interaction (6 articles) is measured by Chu and Tseng (2016) using the following indicators: discussion - if datasets has mechanisms such as forums or feedback; score and rank - if users are able to score or rank the data and the possibility to visualize the number of downloads of the dataset. It has a related dimension called Participation (3 articles) that analyze, according to Zhu and Freeman (2019), criteria such as the possibility to comment, discuss, rate, share and make suggestions to the dataset. When users interact in some way with the database, it generates information that enhances the probability of finding this dataset by search engines. So, discoverability is affected by this group of dimension. Usability is also affected because the participation of the users is a very good opportunity to improve the dataset and make it more suited to the needs of its users. Accessibility is not affected.

Simplicity (5 articles) is a measurement of how simple are the analysis task and the presentation of analysis outputs (Osagie et al., 2017). It has five related dimensions: Presentation (1 article), Functionality (1 article), Friendliness (1 article), Data Utilization (1 article) and Engage-Integrate (1 article). They were all cited only once and the first four appeared in the same article (SP17). In that work the authors decided to divide the evaluation of the data analysis task in more than one dimension where Presentation indicates the type of presentation data, such as tables, images and text; Functionality indicates the degree of completeness of the functionality associated with the dataset; Friendliness indicates whether the interface design is friendly; and Data Utilization indicates whether data-related applications and interfaces are provided. And finally, Engage-Integrate, cited by Zhu and Freeman (2019), verify if the datasets can be downloaded, visualized in other format such as maps, graphics or others, and printed. It is evident that the usability is highly affected by this group of dimension, specially for ordinary users that may have difficult understanding data in a raw format. Discoverability will benefit with the presence of other forms of data representation, because a data interaction tool can generate web traffic that will improve dataset's search capabilities. Accessibility will also be affected, because potential users may have interest in only access data in a graphic format rather than download a file, for example.

Clarity (4 articles) is a dimension whose concerns are the conditions and modalities by which users can obtain, use and interpret data (Stróżyna et al., 2018). Precision (1 article) was found only in Utamachant and Anutariya (2018) and it states that the scope and coverage to the content should be clear. Comprehensiveness (1 article) indicates the richness of the dataset's topics (Wu et al., 2021). Addressability (1 article) description by Kubler et al. (2016) is the extent to which the data publisher provides contact information via 'URL'and 'email'. Discoverability will benefit from a dataset that is clear, precise and filled with important information that can be useful while performing a search. Data usability is also improved because users will comprehend better the meaning of each dataset column and line. Accessibility will not be affected.

Retrievability (3 articles) is, according to Kubler et al. (2018b), the extent to which the described dataset can be retrieved by an agent. Licence-Free (1 article) and Availability (1 article) are the other two related terms and they´re achieved when data is not subject to any limitation on its use due to copyright, patent, trademark or trade secret regulations (Wang and Shepherd, 2020) and when data are available and have attributes that enable them to be retrieved by authorized users and/or applications (Wu et al., 2021), respectively. It is evident that accessibility will be highly affected by this group. Discoverability and usability won´t be influenced.

Consistency (5 articles) dimension is attended by a dataset, according to Li et al. (2018), when the format of a column of data is consistent, such as date format, null expression format, etc. This dimension will affect data usability because null values, for example, are harmful to data analysis and must be correctly treated. Discoverability and accessibility will not be affected.

Duplicity (2 articles) metrics used by Sanabria et al. (2018) are the percentage of record with duplicate records and Uniqueness (2 articles) dimension cited by Li et al. (2018) evaluate if data records are unique and not repeated. As well as null or missing values, duplicate data also needs to be corrected handled, otherwise they can harm data analysis and, consequently, its usability. Discoverability and accessibility will not be affected.

Existence (3 articles) is the extent to which information about dataset´s license, format, size, update frequency, owner is provided (Kubler et al., 2018b). Integrity (1 article) is achieved by a dataset, according to Chu and Tseng (2016), when there are enough data content and metadata. The more metadata you have, better are the chances to find data, improving its

discoverability. Usability is also improved when there are more information about the dataset. Accessibility is not affected.

Conformance (3 articles) is a dimension that analyzes if the information available in the dataset are valid and adhere to a certain format (Neumaier et al., 2016). Normality (1 article), cited only by Wu et al. (2021), indicates whether the metadata values are standardized and consistent. When a dataset contains metadata in a certain standard it will make it easier to be found during a search, therefore affecting its discoverability. Usability is also improved because users can better understand the meaning and the content of a dataset column, for example, if it has a name that is self explanatory or that follows a recognized standard. Accessibility is not affected.

Figure 3 presents the three broad dimensions object of this study (discoverability, accessibility and usability) in a mathematical representation of sets and the 12 group of dimensions placed according to their influence on them. We decided to use only one dimension representing each group in order to do not pollute the image. The group that have more than one dimension, we choose as the representative of the group the dimension that was most cited by the selected articles, or randomly, if they were cited equal times. The groups that affect the three broad dimensions are in the middle of the figure, in the area of intersection of the three sets. When the group affect only two dimensions it is placed in the intersection of those two sets. And if it affect only one dimension it is placed in the area that has no intersection with other set.

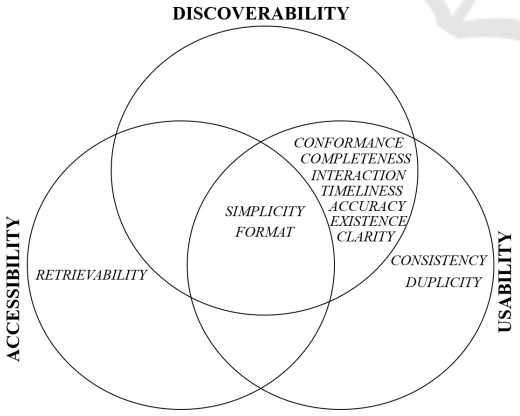

Figure 3: Dimension groups and their influence on the Discoverability, Accessibility and Usability.

Figure 3 shows that only retrievability group do not affect data usability, all of the remaining 11 groups of dimensions have some influence on the usability of the dataset. Most of the dimensions groups (7) influence both discoverability and usability, and 2 dimensions affect discoverability, accessibility and usability.

3.2 SRQ2 - What Kind of Data Sources Are Used in Open Government Data Quality Evaluation Studies?

Most of the articles (12 articles - 57%) did not mention the application domain of the dataset used on the analysis. Healthcare data were the most used (6 articles). Transport, Government and Economic data were analysed by 5 articles each. Datasets including Public Safety, Social and Education information were used by 3 articles each. Environment, Population, Politics, Agriculture, Communication and Culture data appeared only in 2 articles. Energy, Employment, Infrastructure, Religion, Recreation, Housing, Business and Navy data were evaluated by 1 article each.

We believe that the reason because most of the articles did not cite the type of data that were being evaluated was the fact that the main concern of these studies are the evaluation of open government data in general. So, when the researches selected a portal to extract datasets, they´re first concern was not the application domain of those data but its quality aspects according to the chosen dimensions. And that is also reflected on the title and introduction of most of the articles where there is no direct quote to a certain government area. Only two of the selected articles (SP12 and SP17) presented an application domain of interest that guided the selection of the datasets that were analyzed.

3.3 SRQ3 - What Is the Approach Used in Data Analysis?

Human Observation is the approach used by most of the articles (13 articles - 62%). 3 articles chose a Semi-Automatic data analysis type and 4 articles used an Automatic approach. Only 1 article did not mention the way datasets were evaluated.

In the data analysis type that we called human observation, researchers used participants to perform a predefined set of tasks within the dataset and then answer a questionnaire that was used to evaluate data quality. Some works selected ordinary users to participate, others chose users with IT background and there were also a mixed selection of participants. In the semi-automatic approach there were at least one automatic stage in the data quality evaluation process. This automatic stage may be at the data collection step (SP03) or at the quality assessment stage (SP01 and

SP13). An automatic approach is the one who has no human intervention all over the data quality evaluation process, since the data download through the quality metrics analysis of those data and finally aggregating and presenting the results in a friendly form.

3.4 SRQ4 - What Are the Data Quality Non Conformities Found by Researchers?

There were 14 main quality issues reported by the selected articles. Only 2 articles did not mention the problems found in the data quality evaluation. Incomplete Data, Incomplete Metadata, Data Format Not Open, No Interaction Tool were problems reported by 6 articles each. Unstructured Data and No Data Visualization Functionality were found by 5 articles. 4 articles cited No Feedback/Contact Channel as a quality concern. Duplicated Data, Data Access Barriers and No Temporal Information appeared in 3 articles. Missing Values and Lack of Timely Dataset Update was cited by 2 articles. Dataset Fragmentation and Lack of Granularity were quality issues found in only 1 article.

Incomplete data/metadata is related with the dimension group of Completeness which will affect data discoverability and usability. A dataset that is published in a not open format will affect the dimension group of Format and therefore harm data discoverability, accessibility and usability. When dataset has no interaction tool it is prejudicial for the quality dimension of Interaction and will negatively affect data discoverability and usability.

Unstructured data is intimately related with the dimension group represented by Consistency and will affect data usability. The absence of a data visualization functionality may difficult the analysis task affecting the Simplicity dimension which is not good for data discoverability, accessibility and usability.

No feedback/contact channel represents a lack of Interaction, which is a dimension that is directly related with data discoverability and usability.

Duplicated data is quality metric of the dimension Duplicity and can influence data usability. Data access barriers impacts the dimension called Non-Discriminatory, affecting data accessibility. And data with no temporal information is measured by the dimension group represented by Timeliness and will affect data discoverability and usability.

Missing Values affects data Completeness which is related with data discoverability and usability. A lack of timely dataset update affects data quality dimension represented by Timeliness which has a negative influence on data discoverability and usability.

Dataset fragmentation and lack of granularity are a concern of the quality dimension called Primary that affects both data discoverability and usability.

Figure 4 presents the three broad dimensions object of this study (discoverability, accessibility and usability) in a mathematical representation of sets and the data quality non conformities found on the selected articles, placed according to their influence on the 3 broad dimensions.

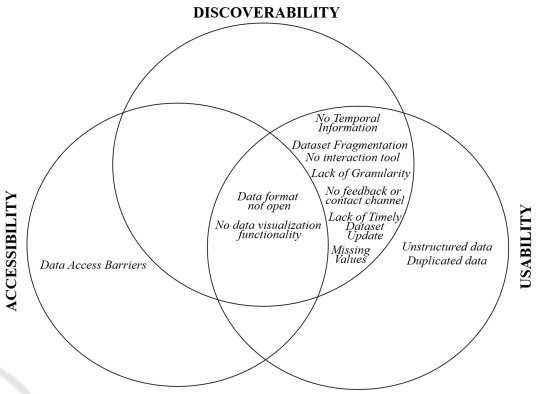

Figure 4: Data quality non conformities and their influence on the Discoverability, Accessibility and Usability.

4 CONCLUSIONS

An effective use of open government data requires minimum quality standards. The evaluation of such requirements is not a trivial task. Several dimensions have been used by authors to evaluate data quality, but three of them cover all the aspects needed for transforming a dataset into value by any stakeholder involved. First, Discoverability, the user must be able to find these dataset on the internet; second, Accessibility, once found it must be accessed without barriers; and finally, Usability, the user must be able to understand and make use of this information.

Even though those three broad dimensions represent all the important aspects of data quality, to evaluate them requires many others dimensions depending on the context or the depth level required by the study. This paper, through a systematic mapping, aimed to describe the different dimensions used by the selected articles to evaluate data quality regarding it´s discoverability, accessibility and usability, and, as already quoted by other authors, we found that there is still no standard or consent about the dimensions that shall be used. And also, some dimensions are used with different nomenclature but with related quality evaluation objectives. For that reason we decided to group these related dimensions and analyse how those dimension groups affect the three broad dimensions object of this

study. This may serve as a tool for a future standardization in the field of data quality evaluation.

The evidence provided by the selected studies also showed that the author´s main concern while doing a data quality evaluation of a dataset is not the application domain of those information but the analysis itself according to the dimensions that are being used. We also noticed that Human Observation is the approach used by most of the articles, and we believe that´s why this approach is effective and easiest than a semi-automatic or automatic data analysis. None of the selected articles reported the use of a machine learning algorithm while analysing the datasets. Data non conformities that were cited was related most with data discoverability and usability. Dimensions related with the form which data is presented to the user (Simplicity) and how the user can interact with the dataset (Interaction) were cited by most of the articles and should be considered while making an open government data available to the public.

There are some threats to validity in our study. First, our research questions may not encompass a full study of the current state of the art of quality evaluation of open government data available on the web. We use the GQM approach to better define the study objective and research questions. It is possible that the search strings we use do not allow the identification of all studies in the area. We mitigate this threat by expanding the number of electronic repositories searched to three. All repositories used are specific of the area of Computing. We cannot guarantee that all relevant primary studies available in electronic repositories have been identified. Some relevant studies may not have been covered by search strings. We mitigate this threat by using alternative search terms and synonyms of major terms in search strings. Each searched electronic repository has its own search process and we don't know how they work or if they work identically. We mitigate this by adapting the search string for each electronic repository and assume that equivalent logical expressions work consistently across all electronic repositories used. The studies were selected according to the defined inclusion, exclusion and quality criteria, but under our judgment. Thus, some studies may have been selected or not selected incorrectly.

REFERENCES

Basili, V. R. and Rombach, H. D. (1988). The tame project: Towards improvement-oriented software environments. *IEEE Transactions on software engineering*, 14(6).

Batini, C., Cappiello, C., and Francalanci (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3).

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

Belhiah, M. and Bounabat, B. (2017). A user-centered model for assessing and improving open government data quality.

Brandusescu, A., Iglesias, C., Robinson, K., Alonso, J. M., Fagan, C., Jellema, A., and Mann, D. (2018). Open data barometer: global report.

Chu, P.-Y. and Tseng, H.-L. (2016). A theoretical framework for evaluating government open data platform. In *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia*.

Clapton, J., Rutter, D., and Sharif, N. (2009). Scie systematic mapping guidance. *London: SCIE*.

Dahbi, K. Y., Lamharhar, H., and Chiadmi, D. (2018). Exploring dimensions influencing the usage of open government data portals. In *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications*.

Dander, V. (2014). How to gain knowledge when data are shared? open government data from a media pedagogical perspective. In *Seminar. net*, volume 10.

Dybå, T. and Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology*, 50(9).

Group, O. G. W. (2007). Eight principles of open government data. https://opengovdata.org/. Accessed in 20-March-2022.

Kassen, M. (2013). A promising phenomenon of open data: A case study of the chicago open data project. *Government information quarterly*, 30(4).

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., and Le Traon, Y. (2018a). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, 35(1).

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., and Le Traon, Y. (2018b). Comparison of metadata quality in open data portals using the analytic hierarchy process. *Government Information Quarterly*, 35(1).

Kubler, S., Robert, Y., Umbrich, J., and Neumaier, S. (2016). Open data portal quality comparison using ahp. In *Proceedings of the 17th international digital government research conference on digital government research*.

Kučera, J., Chlapek, D., and Nečaskỳ, M. (2013a). Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective*. Springer.

Kučera, J., Chlapek, D., and Nečaskỳ, M. (2013b). Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective*. Springer.

Li, X.-T., Zhai, J., Zheng, G.-F., and Yuan, C.-F. (2018). Quality assessment for open government data in china. In *Proceedings of the 2018 10th International Conference on Information Management and Engineering*.

Maali, F., Cyganiak, R., and Peristeras, V. (2010). Enabling interoperability of government data catalogues. In *International Conference on Electronic Government*. Springer.

Máchová, R., Hub, M., and Lnenicka, M. (2018). Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. *Aslib Journal of Information Management*.

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

Magalhães, G. and Roseira, C. (2016). Exploring the barriers in the commercial use of open government data. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS medicine*, 6(7).

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

Nikiforova, A. and McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, 58.

Oliveira, M. I. S., de Oliveira, H. R., Oliveira, L. A., and Lóscio, B. F. (2016). Open government data portals analysis: the brazilian case. In *Proceedings of the 17th international digital government research conference on digital government research*.

Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., and Ojo, A. (2017). Usability evaluation of an open data platform. In *Proceedings of the 18th Annual International Conference on Digital Government Research*.

Sadiq, S. and Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management*, 37(3).

Sanabria, M. A. O., Fernández, F. O. A., and Zabala, M. P. G. (2018). Colombian case study for the analysis of open data government: A data quality approach. In *11th International Conference on Theory and Practice of Electronic Governance*, ICEGOV '18, New York, NY, USA. Association for Computing Machinery.

Stone, P. (2002). Popping the (pico) question in research and evidence-based practice. *Nurs Res*, 15(3).

Stróżyna, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., and Węcel, K. (2018). A framework for the quality-based selection and retrieval of open data-a use case from the maritime domain. *Electronic Markets*, 28(2).

Utamachant, P. and Anutariya, C. (2018). An analysis of high-value datasets: a case study of thailand's open government data. In *2018 15th international joint conference on computer science and software engineering (JCSSE)*. IEEE.

Vetrò, A., Canova, L., Torchiano, M., and Minotas (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, 33(2).

Wang, D., Richards, D., Bilgin, A., and Chen, C. (2021). Advancing open government data portals: a comparative usability evaluation study. *Library Hi Tech*.

Wang, V. and Shepherd, D. (2020). Exploring the extent of openness of open government data–a critique of open government datasets in the uk. *Government Information Quarterly*, 37(1).

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *18th international conference on evaluation and assessment in software engineering*.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Wu, D., Xu, H., Yongyi, W., and Zhu, H. (2021). Quality of government health data in covid-19: definition and testing of an open government health data quality evaluation framework. *Library Hi Tech*.

Yi, M. (2019). Exploring the quality of government open data: Comparison study of the uk, the usa and korea. *The Electronic Library*.

Zhu, X. and Freeman, M. A. (2019). An evaluation of us municipal open data portals: A user interaction framework. *Journal of the Association for Information Science and Technology*, 70(1).

Zuiderwijk, A., Janssen, M., and 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).