INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A framework for open data quality assessment

Abelardo Miguel Ibarra Mendoza

Master in Computer Engineering

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD Bruno Moisés Teixeira Oliveira, Adjunct Professor,
Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Porto

September, 2024

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

A framework for open data quality assessment

Abelardo Miguel Ibarra Mendoza

Master in Computer Engineering

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD Bruno Moisés Teixeira Oliveira, Adjunct Professor,
Escola Superior de Tecnologia e Gestão, Instituto Politécnico do Porto

September, 2024

Information represents to the data what wine represents to the vineyard: the delicious extract and distillate
David Weinberger, *Too Big to Know Write here your dedication*

Acknowledgment

Regrettably, my esteemed parents Abelardo and Nora, who consistently imparted valuable life lessons through their own actions, are no longer present to witness the culmination of their son's accomplishments. Consequently, I wish to express my sincere appreciation to my family for their invaluable support and guidance.

Resumo

Hoje em dia, as entidades públicas e privadas partilham frequentemente os seus dados com os consumidores de dados através de diversas plataformas digitais. Esta partilha de dados, que não tem restrições ou custos, tem sido designada por Dados Abertos. A utilização crescente deste tipo de dados coloca vários desafios, como a qualidade dos dados fornecidos. Garantir a qualidade dos dados é essencial para estabelecer a confiança nos dados abertos e para a sua utilização eficiente em diferentes aplicações.

O principal objetivo desta dissertação é criar uma proposta de uma framework que avalie e meça a qualidade dos Dados Abertos em termos de dimensões e suas métricas. A framework proposta avalia os dados abertos com base em duas extensões distintas. A primeira (Acessibilidade, Interoperabilidade) baseia-se nas caraterísticas sistémicas e técnicas do ecossistema de dados abertos, enquanto a segunda (Completude, Unicidade) incorpora duas propriedades intrinsecamente relacionadas de um conjunto de dados abertos.

O quadro proposto é avaliado utilizando oito conjuntos de dados provenientes de vários domínios que estão acessíveis nos portais da Administração Pública Aberta de Portugal. Os resultados desta dissertação revelam que os actuais conjuntos de dados abertos têm certos problemas de qualidade associados ao sistema e às dimensões técnicas do ecossistema de dados abertos. A contribuição deste trabalho é avaliar os dados abertos dados abertos na perspetiva dos dados em si e dos aspectos técnicos que permitem a sua reutilização.

Palavras Chave: *Dados Abertos*, *Qualidade Dados*, *Dimensões de Qualidade Dados*, *Métricas de Qualidade Dados*

Abstract

Nowadays, public and private entities often share their data with data consumers through various digital platforms. This sharing of data, which has no restrictions or costs, has been called Open Data. The increased use of this kind of data poses several challenges, such as the quality of the data provided. Ensuring the quality of data is essential for establishing trust in open data and for its efficient utilization in different applications.

The main goal of this dissertation is to create a proposal for a framework that assesses and measures the quality of Open Data in terms of dimensions and its metrics. The proposed framework evaluates open data based on two distinct extend. The first one (Accessibility, Interoperability) is founded upon the systemic and technical characteristics of the open data ecosystem, whereas the second extends (Completeness, Uniqueness) incorporates two inherently related properties of an open dataset.

The proposed framework is assessed using eight datasets originating from various domains that are accessible on the Portuguese Open Government portals. The findings of this dissertation reveal that open datasets today have certain quality issues associated with the system and technical dimensions of the open data ecosystem. The contribution of this work is to assess open data from the perspective of the data itself and technical aspects to allow their reuse.

Keywords: *Open Data*, *Data Quality*, *Data Quality Dimensions*, *Data Quality Metrics*

Contents

List of Figures

List of Tables

List of Acronyms

AI: Artificial Intelligence

DQ: Data Quality

GQM: Goal Question Metric

LD: Linked Data

LOD: Linked Open Data

OD: Open Data

ODP: Open Data Portals

OGD: Open Government Data

OGDP: Open Government Data Portals

UML: Unified Modeling Language

CHAPTER 1

Introduction

Nowadays, data is regarded as one of the most crucial resources for governments, companies, and individuals [26], [73]. Each day, a vast amount of data is generated from various sources [69], [16]. Public sector organizations around the globe are putting into action Open Data initiatives, with the expectation that these initiatives will encourage economic expansion, enhance transparency and responsibility, and facilitate better interaction between data users (usually citizens) and data providers [52]. While Open Data is not only for government use, the vast majority of datasets published on Open Data portals come from public bodies. As a result, the primary focus of Open Data is usually on open government data. In recent years, there has been a notable rise in the number of cities making open datasets available to the general public [26].

Gurin [21] defines Open Data (OD) as accessible public data that individuals and organizations can use to develop new ventures, discover patterns or trends, make data-driven decisions, and answer complex problems. Similarly, the Open Data Handbook [50] defines OD as "*data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike*". OD is a form of information that is produced, collected, or exchanged. It enables organizations to enhance their performance and can have both economic and social impacts [19].

However, simply compliance with these principles and utilizing the data itself does not ensure its quality; it must also adhere to established regulations [73], and standards [23]. OD should be accompanied by data discovery mechanisms and other activities to provide indications of data quality and reliability [26].

The importance of Data Quality (DQ) resides in its capacity to establish the basis for the degree of confidence one can invest in the conclusions derived from evaluation. DQ specifically pertains to the cleanliness of the data (absence of errors and gaps), untainted (lack of bias), and the consistency of the data (minimal discrepancies) [26]. OD is frequently utilized in order to make crucial decisions based on the findings of its analysis [39]. The quality of OD has a significant influence on the decision-making process [45]. This dissertation aims to explore the description of OD quality, the considerations that could assist in developing the comprehensive data quality framework, and its potential application in assessing the quality of OD.

1.1. Motivation

Data is experiencing exponential growth rates, and governments and public and private institutions require data daily to provide better services and have a positive economic impact. The continuous advancement of technologies, which is greatly impacted by the growth of the Internet, serves as the main catalyst for this phenomenon. This progression has resulted in a rise in the creation of data and a subsequent need for the availability of information in formats that can be used again, all thanks to the enhanced capabilities of information systems. In recent years, OD has emerged within this context.

The utilization of OD can bring numerous benefits to stakeholders [65]. OD has the potential to revolutionize every sector of the economy and generate over $3 trillion in global economic value each year [12]. However, the success of OD may be threatened if there are issues with its quality. The lack of quality of published data is a barrier to the reuse process becoming massive. OD frequently presents issues with data quality that need to be identified prior to utilizing the data for analysis, as any deficiencies in data quality could result in substantial losses [47].

OD is regularly generated by government organizations, however, organizations in the private sector are acknowledged as sources of OD as well. Some businesses choose to make data publicly available as a means of promoting transparency and innovation [8]. These datasets cover a broad spectrum of topics, including traffic, weather, geography, tourism information, statistics, business, census, budget and spending, education quality, and so on. They are available through Open Government Data Portals (OGDP) [63]. Over time, various initiatives have been dedicated to enhancing the accessibility, usefulness, and compatibility of the data on OGDP.

Initiatives such as 5 Stars data[1], propose a hierarchical system of multiple levels of usefulness in OD. More specific frameworks have emerged to assess the quality of data, with a particular emphasis on metadata [29]. However, the majority of these frameworks fail to address a fundamental issue: errors in the data values themselves [59]. As users have started incorporating open data into their apps, they are bringing to the attention quality problems with these datasets[2], which the data publishers have since fixed [14]. The quality of data becomes the main concern due to the fact that OD can encompass a wide range of information and come from various data sources [22].

One common misunderstanding about OD is the belief that merely granting access to data is enough for it to be effectively reused [24]. It is imperative that the data can be trusted and deemed reliable. In order to achieve this, the quality of OD can be assessed based on relevant dimensions and metrics.

This serves as a principal motivation to consider the necessary aspects to assist in

[1] http://5stardata.info/.

[2] http://bit.ly/opendata-betterdata

the developing a solution for managing data quality, as well as its potential application in evaluating the quality of open data.

1.2. Objective

In the context of Open Data, due to the quality data issues, arises the necessity for a framework that can assess the quality of the data being published. Even if the data is of good quality, it may not be valuable (or add value) to users because it is often used for a purpose that was not initially envisaged during the data collection process, and therefore, it may not be the data they need to make decisions [58]. This evaluation should guarantee that the data is valuable to those who wish to utilize it.

The objective of data quality research for OD is to understand the challenges and issues related to the data quality in order to identify necessary factors for developing a framework for assessing the data quality of datasets.

1.3. Research Questions

The dissertation aims to address the research questions below in order to accomplish the proposed objectives:

(1) What are the commonly used quality dimensions and metrics to assess the quality of Open Data?
(2) What are the frameworks and methods that are used to evaluate the quality of Open Data?
(3) What are the current limitations in technology for assessing the quality of Open Data?
(4) What are the main tools used to evaluate the quality of Open Data and what limitations are associated with them?
(5) Can data quality issues be identified in Open (Government) Data by applying a metric-based assessment system that depends on technological and intrinsic data dimensions?

CHAPTER 2

State Of The Art

This chapter explores the present state of data quality within the realm of Open Data. It provides a thorough examination of the existing literature by utilizing a systematic search approach and inclusion criteria based on the PRISMA[1] methodology. The search involved academic data repositories such as Scopus[2] and Google Scholar[3], resulting in 224 papers. After a careful selection process, 66 articles were considered relevant to this study. These chosen papers, published between 2011 and 2023, encompass a range of topics based on the research questions. The goal of this chapter is to develop a comprehensive understanding of the current state of the art through an analysis of these papers.

2.1. Systematic Review

A search would need to be conducted in a database to discover the work pertaining to the dissertation. It should be noted that Scopus served as the primary resource for research, while Google Scholar was utilized as a secondary source to remain informed about the works connected to the dissertation.

The method of inquiry employed in these two databases varied. In Scopus, a search was carried out by means of a query, whereas in Google Scholar, searches were conducted using specific keywords as needed.

2.2. Identification of keywords and search expressions

In order to comprehend the latest developments, the goal is to address the question: "What is the state of the art of data quality in the contex of Open Data?". For the selection of the study, the first step involves identifying keywords and formulating a search string to explore the databases. A total of 9 keywords were defined. By utilizing keywords and leveraging the 'AND' and 'OR' connectors, two search expressions (Table 2.1) were crafted to choose the pertinent articles. One search string is designed to retrieve literature from SCOPUS, while the other is intended for accessing relevant works from Google Scholar.

In addition to the mentioned database search method, a filtering technique was also employed to select the most precise works, given the extensive number of results acquired. Criteria for Inclusion and Exclusion were set up (Table 2.2).

Considering the manner in which the search was conducted and the filters that

[1]https://www.prisma-statement.org/
[2]https://www.scopus.com/
[3]https://scholar.google.com/

Table 2.1. Keywords and Search Expressions

Keywords	*Data Quality, Data Quality Dimensions, Data Profiling, Data Quality Metrics, Data Quality Assurance, Data Quality Score, Data Quality Monitoring, Data Observability, Data Trust, Open Data*	Results
Scopus - principal query	*(TITLE-ABS-KEY ("Data Quality") OR TITLE-ABS-KEY ("Data Quality Dimensions") OR TITLE-ABS-KEY ("Data Profiling") OR TITLE-ABS-KEY ("Data Quality Metrics") OR TITLE-ABS-KEY ("Data Quality Assurance") OR TITLE-ABS-KEY ("Data Quality Score") OR TITLE-ABS-KEY ("Data Quality Monitoring") OR TITLE-ABS-KEY ("Data Observability") OR TITLE-ABS-KEY ("Data Trust") AND TITLE-ABS-KEY (Open Data))*	560
Google Scholar - principal query	*allintitle: "Data Quality" OR "Data Quality Dimensions" OR "Data Profiling" OR "Data Quality Metrics" OR "Data Quality Assurance" OR "Data Quality Score" OR "Data Quality Monitoring" OR "Data Observability" OR "Data Trust" "Open Data"*	247

Table 2.2. Exclusion and inclusion filters in the search for related work.

Inclusion	Exclusion
Articles	Not being Articles
Written in English,Spanish and Portuguese	Articles written in another language
From 2010 to 2023	Older than 2010
In the area of Computer Science	Not in the area of Computer Science
In the area of Engineering	Not in the area of Engineering
Open Access or ISCTE subscription	Paid works

were applied, the phase of analyzing and selecting the results commenced. A methodology known as PRISMA[4] (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) was implemented. Additionally, during the search process, significant references were discovered in the analyzed works. It should be mentioned that the restrictions listed in Table 2.2 did not apply to these references.

We started with 807 (Table 2.1) relevant studies. Following the application of inclusion and exclusion criteria (query filters) the results are shown in Tables 2.3, 2.4, 2.5..

Table 2.3. Query Results - First Filter

Source	Query filter	Results
Scopus	*(TITLE-ABS-KEY ("Data Quality") OR TITLE-ABS-KEY ("Data Quality Dimensions") OR TITLE-ABS-KEY ("Data Profiling") OR TITLE-ABS-KEY ("Data Quality Metrics") OR TITLE-ABS-KEY ("Data Quality Assurance") OR TITLE-ABS-KEY ("Data Quality Score") OR TITLE-ABS-KEY ("Data Quality Monitoring") OR TITLE-ABS-KEY ("Data Observability") OR TITLE-ABS-KEY ("Data Trust") AND TITLE-ABS-KEY (Open Data)) AND (LIMIT-TO (SUBJAREA , "COMP"))*	390
Google Scholar	*allintitle: "Data Quality" OR "Data Quality Dimensions" OR "Data Profiling" OR "Data Quality Metrics" OR "Data Quality Assurance" OR "Data Quality Score" OR "Data Quality Monitoring" OR "Data Observability" OR "Data Trust" "Open Data" -"big data" -"linked open data" AND Publication Date between 2010 and 2023*	198

[4]https://www.prisma-statement.org/

Table 2.4. Query Results - Second Filter

Source	Query filter	Results
Scopus	*(TITLE-ABS-KEY ("Data Quality") OR TITLE-ABS-KEY ("Data Quality Dimensions") OR TITLE-ABS-KEY ("Data Profiling") OR TITLE-ABS-KEY ("Data Quality Metrics") OR TITLE-ABS-KEY ("Data Quality Assurance") OR TITLE-ABS-KEY ("Data Quality Score") OR TITLE-ABS-KEY ("Data Quality Monitoring") OR TITLE-ABS-KEY ("Data Observability") OR TITLE-ABS-KEY ("Data Trust") AND TITLE-ABS-KEY (Open Data)) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "bk") OR LIMIT-TO (DOCTYPE , "ch"))*	128
Google Scholar	*allintitle: "Data Quality" OR "Data Quality Dimensions" OR "Data Profiling" OR "Data Quality Metrics" OR "Data Quality Assurance" OR "Data Quality Score" OR "Data Quality Monitoring" OR "Data Observability" OR "Data Trust" "Open Data" -"big data" -"linked open data" -"Open Government Data" AND Publication Date between 2010 and 2023 AND Language "English, Portuguese,Spanish "*	121

Table 2.5. Query Results - Third Filter

Source	Query filter	Results
Scopus	*(TITLE-ABS-KEY ("Data Quality") OR TITLE-ABS-KEY ("Data Quality Dimensions") OR TITLE-ABS-KEY ("Data Profiling") OR TITLE-ABS-KEY ("Data Quality Metrics") OR TITLE-ABS-KEY ("Data Quality Assurance") OR TITLE-ABS-KEY ("Data Quality Score") OR TITLE-ABS-KEY ("Data Quality Monitoring") OR TITLE-ABS-KEY ("Data Observability") OR TITLE-ABS-KEY ("Data Trust") AND TITLE-ABS-KEY (Open Data)) AND PUBYEAR > 2010 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA,"COMP")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"bk") OR LIMIT-TO (DOCTYPE,"ch")) AND (LIMIT-TO (LANGUAGE,"Spanish") OR LIMIT-TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"Portuguese")) AND (LIMIT-TO (EXACTKEYWORD,"Open Data") OR LIMIT-TO (EXACTKEYWORD,"Data Quality") OR LIMIT-TO (EXACTKEYWORD,"Open Datum") OR LIMIT-TO (EXACTKEYWORD,"Data Completeness") OR LIMIT-TO (EXACTKEYWORD,"Data Cleaning") OR LIMIT-TO (EXACTKEYWORD,"Quality Of Data") OR LIMIT-TO (EXACTKEYWORD,"Completeness") OR LIMIT-TO (EXACTKEYWORD,"Data Accuracy") OR LIMIT-TO (EXACTKEYWORD,"Data Assessment") OR LIMIT-TO (EXACTKEYWORD,"Data Curation") OR LIMIT-TO (EXACTKEYWORD,"Data Profiling") OR LIMIT-TO (EXACTKEYWORD,"Quality Issues") OR LIMIT-TO (EXACTKEYWORD,"Data Set") OR LIMIT-TO (EXACTKEYWORD,"Data Quality Assessment"))*	103
Google Scholar	*allintitle: "Data Quality" OR "Data Quality Dimensions" OR "Data Profiling" OR "Data Quality Metrics" OR "Data Quality Assurance" OR "Data Quality Score" OR "Data Quality Monitoring" OR "Data Observability" OR "Data Trust" "Open Data" -"big data" -"linked open data" -"Open Government Data" AND Publication Date between 2010 and 2023 AND Language "English, Portuguese,Spanish "*	121

The number of related studies was reduced to 224 as detailed in the Table 2.6.

Table 2.6. Summary of search results by information source.

Source	Principal Query (Table 2.1)	First Filter (Table 2.3)	Second Filter (Table 2.4)	Third Filter (Table 2.5)
Scopus	560	390	128	103
Google Scholar	247	198	121	121
Total	**807**	588	249	**224**

The next step was to analyse the title and abstract of the papers found, 5 duplicate documents were identified between the two data sources and subsequently removed, at the end of this step there were 153 papers left. As there were still too many documents to perform a complete analysis of each one, a skimming read was performed on each work, and in the end, 66 papers were selected. These documents were subsequently imported to Mendeley Reference Manager[5]. In the end 66 works where fully analyzed and used. This process is summarized in the Table 2.7. The final list of selected papers is presented in Appendix A.

Table 2.7. Filtering process of related works.

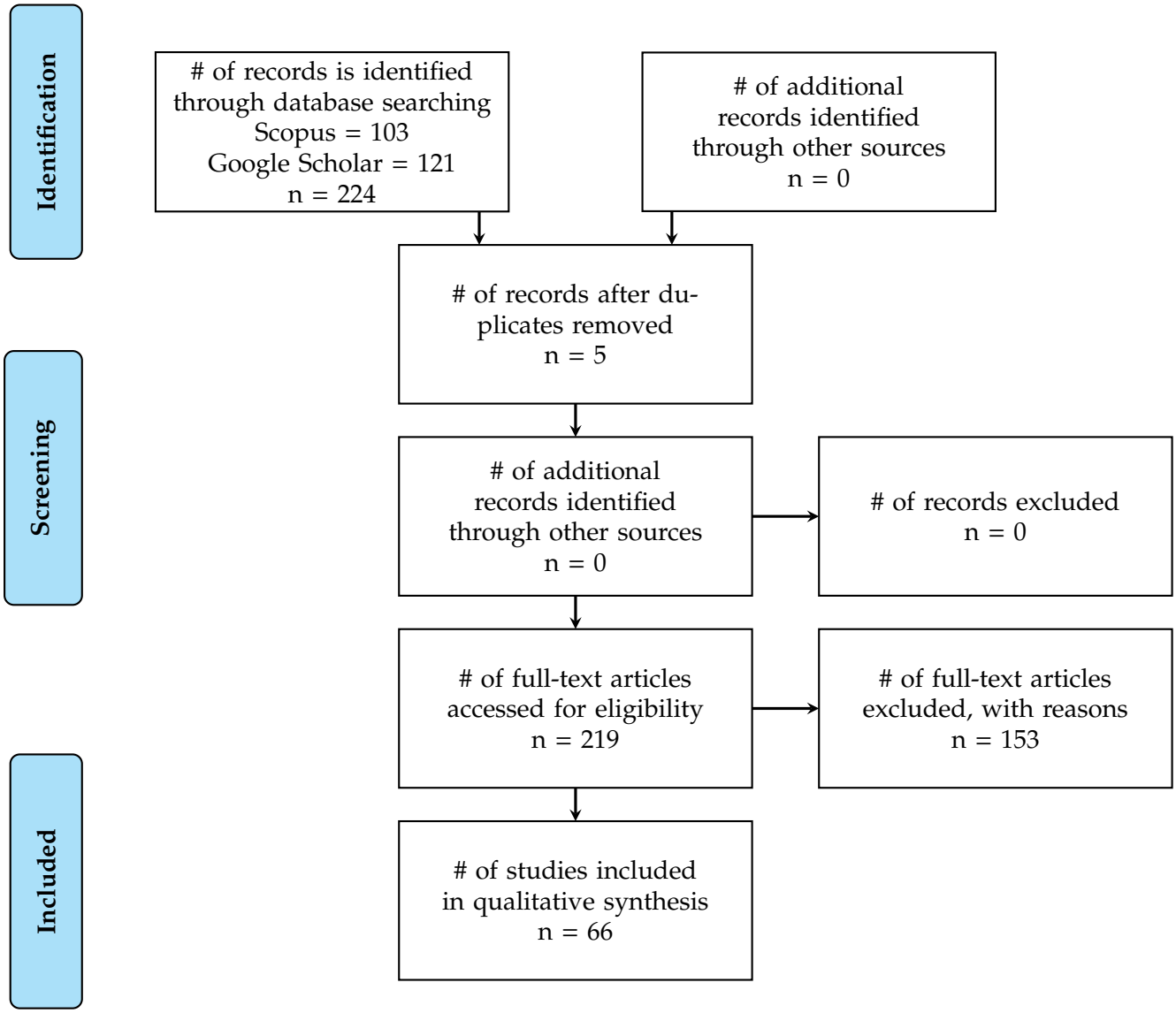
Inclusion	**Nº**
Initial search with the inclusion and exclusion criteria	224
Duplicate documents	5
Analysis of the title and abstract of the documents	219
Skimming of documents	153
Complete document analysis and utilisation	66

The description above was based on the PRISMA methodology, and the respective workflow can be found in Table 2.8.

By taking this factor and the other analyses of articles that have already been discussed into consideration, the composition of the state-of-the-art was accomplished.

[5]https://www.mendeley.com/

Table 2.8. PRISMA - Flow diagram of conducted systematic literature review



2.3. Related Work

2.3.1. Data Quality

Data Quality (DQ) encompasses various aspects and can be interpreted in numerous ways from different viewpoints [71]. The concept of "Data quality" is defined in various ways. Nevertheless, data is typically considered to be of high quality when it is suitable for its intended purposes [46]. It is crucial in both decision-making and operational procedures [4]. DQ is described as a crucial matter in the management of data.

Loshin [34, p. 101], provide a simple definition of DQ : *"fitness for use"*. Emphasizes that the initial step in classifying requirements and measurement goals for data is to define a set of dimensions outlined in Table 2.9. He focuses first on data structure and describes characteristics of high-quality data models, afterward, attributes of data values.

Table 2.9. Loshin's dimension of data quality

Group	Dimension	Definition
Data models	Clarity of Definition	Naming convention that is given to tables, fields, and relations in the system.
	Comprehensiveness	Scope that encompasses all the relevant information.
	Flexibility	Reflects the capacity to change in reactions to new requirements.
	Robustness	Changes in the modeled world without excessive changes to the data model.
	Essentialness	Should not contain any additional information, unless it is necessary.
	Granularity	Refers to the quantity of objects that are employed to symbolize a single concept.
	Precision of Domains	Pertains to the level of detail that can be assigned to the value of an attribute..
	Homogeneity	Data model is designed to support a specific category of entities, model gradually expands to include multiple categories of entities within the same database.
	Naturalness	Every represented attribute should correspond to a natural object in the world being modeled.
	Identifiability	Each entity type must allow for a unique identification(primary key).
	Obtainability	Determines if the information that needs to be modeled can be gathered and saved.
	Relevance	If eliminated the attribute, would have no impact on the current or future situation.
	Simplicity	Refers to the complexity in the data model.
	Semantic Consistency	Meanings and the names of objects within the data set are consistent.
	Structural Consistency	Representation of similar attribute values, both within the same data set.
Data Values	Accuracy	Data values agree with an identified source of correct information.
	Null Values	Null value can be understood as a value that is absent, not necessarily.
	Completeness	Certain attributes in a data set are anticipated to possess assigned values.
	Consistency	Data values in one data set being consistent with values in another data set.
	Currency/Timeliness	The degree to which information is current with the world that it models.
Data Domains	Enterprise Agreement of Usage	Conformity to the usage of the enterprise data domain of record instead of relying on their own data set.
	Stewardship	Responsibility has been assigned for the stewardship of information domains
	Ubiquity	Degree to which different departments in an organization use shared reference data.
Data Presentation	Appropriateness	Format and presentation of the data match users needs.
	Correct interpretation	Provides the user with everything required for the good interpretation of information.
	Flexibility	Describes ability of the system to adapt to changes in represented information.
	Format Precision	Precision of the presentation of an attribute's value.
	Portability	In heterogeneous environments, interoperability the presentation of data is familiar to the users.
	Representation Consistency	Instances of data are represented in a format that is consistent with the domain.
	Representation of Null Values	Recognizable form for presenting that null value that does not conflict with valid values.
	Use of Storage	How effectively the storage requirements are offset by other needs.
Information Policy	Accessibility	Degree of ease of access to the information.
	Metadata	Is data about the data in the system.
	Privacy	Involves the ways unauthorized users are prevented from accessing data.
	Redundancy	Storage of multiple copies of equivalent data values.
	Security	Protection of data from harm, unauthorized modifications, or unwanted destruction.
	Unit Cost	Costs incurred to obtain values, maintain levels of data quality, store data, and so on.

Source: Adapted from [34]

According to Kitchin [26], the quality of data is determined by how clean (free from errors and gaps), untainted (free from bias), and consistent (with few discrepancies) the data is.

Olson [49, p. 24] defines DQ as follows: *"data has quality if it satisfies the requirements of its intended use. It lack quality to the extent that it does not satisfy the requirement"*, emphasizes the importance of data accuracy. In order to satisfied the intended use, recommends evaluate six data quality dimensions (aspects): accuracy, timeliness, relevance, completeness, understood and trust. Sebastian-Coleman [61, p. 39] also states that *"data quality is about whether data meets implicit or explicit expectations of people who will use the data. How someone judges the quality of data depends on what that person expects from the data"*. Therefore, the quality of data is directly connected to the intended or predetermined objectives of the data.

The term "dimension" is employed to denote the various aspects of data that can be measured and through which the quality of the data can be described and quantified. Its synonyms include component, indicator, characteristic, factor, criteria, category, issue, phase, theme, and functionality [65]. Actually, there is not consensus on which dimensions are the core for DQ measurement, dimensions explored include completeness, validity, timeliness, consistency, and integrity [61]. The diversity of the lists that describe the dimensions of data quality is very extensive. Southekal [67, p. 39] states that: *"The word 'dimension' is used to identify aspects of data elements that can be defined, quantified, measured, implemented, and tracked."*

Bicevskis [44], presents the viewpoint that DQ refers to how well a specific dataset and its attributes align with a specific purpose or need, which is determined by the data consumer. It may be necessary to establish distinct data quality criteria for the same data, depending on the specific use case.

The standard ISO/IEC 25012:2008 [23] provides a comprehensive quality model for data that is presented in a structured format within a computer system. This model defines data quality as *"the degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions"* related to inherent perspective, and *"degree to which data quality is accessed and preserved through a computer system when the data is used under specific conditions"*, related to system-dependent one. The inherent and system-dependent data quality characteristics (dimensions in the following of this document) defined in the ISO standard are described in Table 2.10.

Artificial Intelligence (AI) has achieved success in various applications, but one area where its impact may not be as obvious is in improving data management, particularly in terms of quality. AI tools are capable of examining data to identify prohibited values. Some of these values are automatically corrected, while others are assigned to an individual or a group for correction [15].

Table 2.10. ISO/IEC 25012 data quality characteristics

Characteristic	Definition	Inherent	System-Dependent
Accuracy	The degree to which the data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.	✓	
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.	✓	
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.	✓	
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.	✓	
Currentness	The degree to which data has attributes that are of the right age in a specific context of use.	✓	
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.	✓	✓
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.	✓	✓
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use. Confidentiality is an aspect of information security (together with availability, integrity) as defined in ISO/IEC 13335-1:2004.	✓	✓
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.	✓	✓
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.	✓	✓
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.	✓	✓
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.Some information about data understandability are provided by metadata.	✓	✓
Availability	The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.		✓
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.		✓
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.		✓

Source: Adapted from [23]

The standard ISO/IEC 5259-2:2024 Artificial Intelligence - Data Quality for Analysis and Machine Learning (ML) - Part 2: Data Quality Measures [6] describe a data quality model by defining data quality metrics and characteristics using ISO/IEC 25012 and ISO/IEC 25024 as a guide. Terms and definitions such as raw data, data frame, data value, entity, and so on are included. Issues in data integrity, including but not limited to incomplete, erroneous, or obsolete data, have the potential to unfavorably influence analytics and ML processes and outcomes.

It is worth mentioning that this standard also includes some new quality characteristics. The primary additional characteristics that should be considered for AI are those that pertain to groups of data, specifically datasets, rather than individual pieces of data. This standard is also in the process of defining various measures of datasets quality, including dataset representativeness. Additionally, they should incorporate

[6] https://www.iso.org/standard/81860.html

some characteristics specific to the dataset. When using data, whether it is single or grouped, it is also necessary to consider another standard that addresses data quality aspects related to management [41].

In recent decades, technology has made significant advances, public and private initiatives for share data where implemented, permitted that a novel idea originating from the concept of "data" has arisen - "open data", which presents new challenges due to its inherent nature. This subject will be further explored in the following subsection.

2.3.2. Open Data

Due to the exponential growth of the amount of data available in great number of information sources around the world, efforts have proliferated to integrate and classify data, enabling its publication and subsequent analysis, whether for private purposes or for social benefit. It is not just organizations that generate data, but also individuals who utilize electronic devices in their daily activities, as well as machines that are linked via the Internet of Things (IoT) [62]. The volume of data will keep increasing. These factors make it necessary to manage the data and their quality [60]. Within this expansive context, there is a powerful movement that has embraced the concept of Open Data [40]. Open data may be advantageous to academic institutions, individual researchers, and society as a whole [33]. According to Immonen et al. [22], OD can be categorized into three types: structured data, which follows a strict data model, semi-structured data, which adheres to an evolving data model, and unstructured data, which is not linked to any specific data model.

2.3.2.1. History

According to Thompson [69], the Open Data movement has mainly occurred in developed countries with a focus on federal and national data. In addition, the majority of US states have implemented Open Data laws and regulations.

One of the primary forces behind the movement is Tim Berners-Lee, widely regarded as the founding figure of the World Wide Web. According to Monino and Sedkaoui [40, p. 27], Tim Berners-Lee observes on the subject of Open Data that *"If we share data online – public data, scientific data, citizens' data, whatever – then other people will be able to develop marvelous creations from that data that we could never even have imagined"*. The primary objectives include making local, regional, and national data accessible to the public in electronic form.

Thompson [69] states *"Personal identifiable information (PII) will not be designated as Open Data, but data such as transactions, travel, and other data are perfect candidates to be marked as Open Data"*. Another significant factor contributing to the Open Government Data (OGD) movement is the "Memorandum on Transparency and Open Government," which was signed by US President Barack Obama shortly after he took office in January 2009 [11, p. 5]. The goal was to create a modern partnership among politicians, public administration, industry, and private individuals by promoting increased transparency,

democracy, participation, and cooperation. In European nations, Open Government is frequently seen as a complementary aspect to e-government [55].

In terms of terminology, there are various acronyms that vary from one another in the literature. While the term Open Government Data (OGD) is referred to public data produced by public sector bodies, the abbreviation "Open Data" is also used, it includes whatever data such as: government, businesses, health, insurances, mappings, among others [55]. OGD utilization is still minimal even with the abundance of datasets available [76].

The concept of Open Data, which involves making government data accessible to the public, spread rapidly in the United States before gaining popularity in Germany, France, and various other nations [40]. It is important to mention that OD has been utilized since the 1970s[7]. It is feasible to condense the chronology of Open Data's progress in the subsequent timeline in the Figure 2.1.

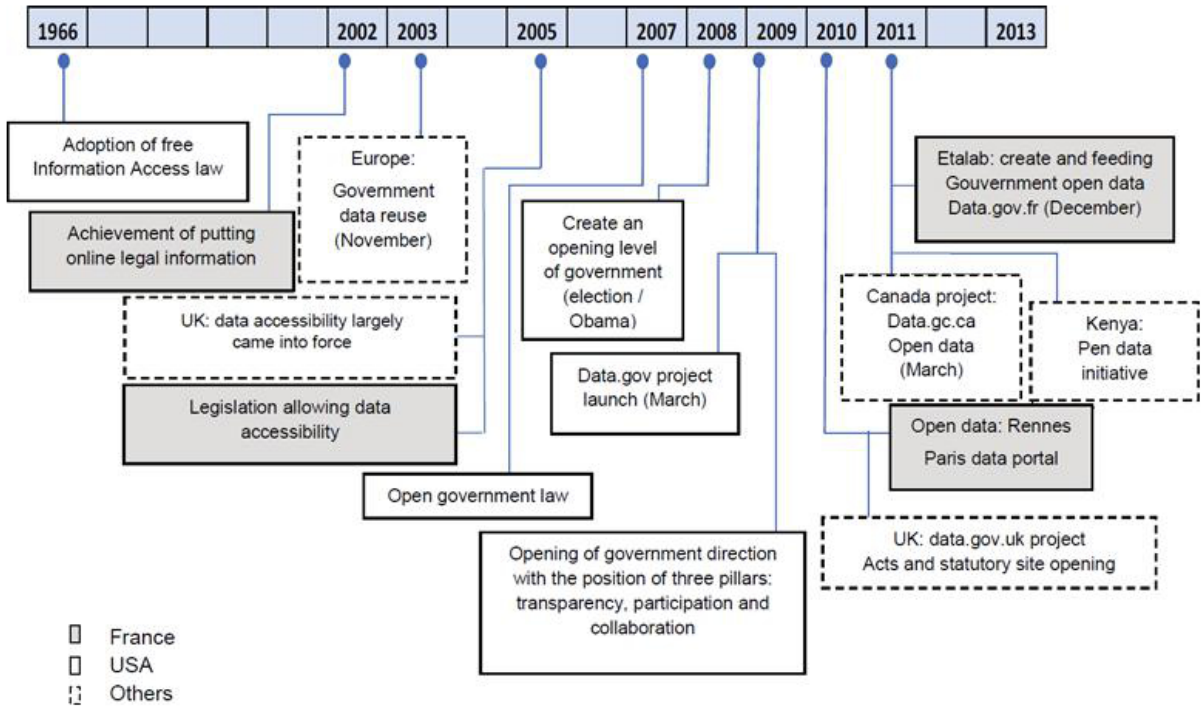


Figure 2.1. Open Data history
Source : Monino and Sedkaoui [40, p. 25]

According to Monino and Sedkaoui [40], OD refers to digital data that can be either private or public. This data is generated by collective entities or public services. It is distributed in a structured manner following a specific method, along with an open license that ensures unrestricted access to it. Furthermore, this OD can be reused by anyone without encountering any technical, legal, or financial limitations. As pointed out by Koltay [27], data reuse is the use of data by someone other than the one that originally collected it. OD consists of various sources and types of data [40]:

[7]https://en.wikipedia.org/wiki/Open_data_in_the_United_States

- Public data or information originating from the public sector. This encompasses all data gathered by public entities at every level;
- Data from scientific research, especially from research funded by the public;
- Data from the private sector can be disclosed to the public by providing appropriate incentives and ensuring privacy protections.

Sadiq and Indulska [58, p. 150] states *"Open Data is data made freely available by governments, organizations, researchers, among others, for use by anyone without copyright restrictions"*, OD includes two basic features: the data must be publicly available for anyone to use, and it must be licensed in a way that allows for its reuse. OD should also be relatively easy to use, although there are certain gradations of openness of such data. There is general agreement that OD should be available free of charge or at minimal cost.

According to Campos Zabala [8], when utilizing OD, it is essential to understand the licensing and legal aspects linked to each dataset. The licenses establish the conditions for using, altering, and distributing the data. Licenses define the terms under which the data can be used, modified, and shared. Some common open data licenses include the Creative Commons (CC)[8] licenses and the Open Data Commons licenses. An outline of Licensing types for CC is provided in the Table 2.11.

Table 2.11. Licensing types for Creative Commons (CC)

License	Usage rights	Modification rights	Sharing rights
CC0 (Creative Commons Zero)	Unrestricted use	Unrestricted modification	Unrestricted sharing
CC BY (Attribution)	Use with attribution	Modification with attribution	Sharing with attribution
CC BY-SA (Attribution-ShareAlike)	Use with attribution	Modification with attribution	Sharing with attribution, under the same license
CC BY-ND (Attribution-NoDerivs)	Use with attribution	No modification	Sharing with attribution
CC BY-NC (Attribution-NonCommercial)	Use with attribution for non-commercial purposes	Modification with attribution for non-commercial purposes	Sharing with attribution for non-commercial purposes
CC-BY-NC-SA (Attribution-NonCommercial-ShareAlike)	Use with attribution for non-commercial purposes	Modification with attribution for non-commercial purposes	Sharing with attribution, under the same license, for non-commercial purposes
CC-BY-NC-ND (Attribution-NonCommercial-NoDerivs)	Use with attribution for non-commercial purposes	No modification	Sharing with attribution for non-commercial purposes

Source: Adapted from Campos Zabala [8]

Batini and Scannapieco [4] suggest that OD are freely available machine-readable data. The philosophy behind OD has been long established in public bodies, but the term "Open Data" itself is recent, gaining popularity with the rise of the Internet and World Wide Web and, especially, with the launch of Open Data Government initiatives.

Personal data about individuals, identity information, data about national security, data related to the military, and data that would disadvantage individuals or groups

[8]https://creativecommons.org/share-your-work/cclicenses/

are all excluded from Open Data provided by governments [69].

The Open Knowledge Foundation (OKF)[9] outlines principles that establish the concept of "openness" concerning data and content : *"Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness*)". Expressed in a concise manner, *"Open data and content can be freely used, modified, and shared by anyone for any purpose"* .

OpenGovData[10] has established nine principles of OD as detailed in Table 2.12.

Table 2.12. OpenGovData's principles of open data

Data must be completed	All data are made available, subject to statutes of privacy, security or privilege limitations.
Data must be primary	Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.
Data must be timely	Data is made available as quickly as necessary to preserve the value of the data.
Data must be accessible	Data is available to the widest range of users for the widest range of purposes.
Data must be machine-processable	Data is reasonably structured to allow automated processing.
Access must be non-discriminatory	Data is available to anyone, with no requirement of registration.
Data formats must be non-proprietary	Data is available in a format over which no entity has exclusive control.
Data must be license-free	Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed as governed by other statutes.
Compliance must be reviewable	A contact person must be designated to respond to people trying to use the data or complains about violations of the principles and another body must have the jurisdiction to determine if the principles have been applied appropiately.

Source: Adapted from http://opengovdata.org

Gartner[11], the worldwide technology consulting company, offers the following definition of Open Data: *"Open data is information or content made freely available to use and redistribute, subject only to the requirement to attribute it to the source. The term also may be used more casually to describe any data that is shared outside the organization and beyond its original intended use, for example, with business partners, customers or industry associations"*.

Some OD is also referred to as Linked Open Data (LOD) and is based on the concept that the current mechanisms used for sharing and connecting documents on the Internet can also be utilized for sharing and connecting data and metadata related to these documents (Table 2.13).

Table 2.13. Juxtaposition of the concepts of Open Data, Linked Data, and Linked Open Data.

Representation\degree of openness	**Possibly closed**	**Open (cf. opendefinition.org)**
Structured data model (i.e. XML, CSV, SQL etc.)	**Data**	**Open Data**
RDF data model (published as Linked Data)	**Linked Data**(LD)	**Linked Open Data**(LOD)

Source : Adopted from Auer [3]

[9]https://opendefinition.org/
[10]https://https://opengovdata.org/
[11]https://www.gartner.com/en/information-technology/glossary/open-data

In 2006, Tim Berners-Lee introduced the phrase Linked Data (LD) to describe a collection of guidelines for sharing and connecting organized data on the Internet. Thus, LD is define by the use of the four rules:

(1) Use URIs as names for things
(2) Use HTTP URIs, so that these URIs can refer to those things
(3) Provide useful information at the URI's destination (including the use of standards, e.g. RDF, SPARQL)
(4) Include links to other URIs.

Berners-Lee [7] states that LOD refers to LD that is made available under an open license, allowing for its free reuse without any restrictions.

Monino and Sedkaoui [40, p. 27]) state that Open Data should possess three key attributes:

- **Technical**
  The raw data should be easily usable in an automated way and should be available in open-source formats whenever possible;
- **Legal**
  The licenses must clearly define the rights and obligations of both data owners and individuals interested in reusing the data. These licenses should aim to be as open as they can be;
- **Economic**
  There should be minimal or no royalties in order to avoid any potential barriers to reusing the content.

2.3.2.2. Open Data Portals

OD refers to data that is made available on websites that can be accessed by the public. The purpose of making this data accessible is to allow individuals or organizations to reuse and distribute it without any limitations or restrictions.

OD is typically accessible via Open Government Data Portals (OGDP) that facilitate analysis in order to empower informed decision-making. The OGDP serves as a centralized national platform where public sector entities within the nation share their data with the purpose to enhance transparency and integrity [55]. In addition, the OGDP is a simple website interface through which they facilitate the use of published data so that citizens and other non-governmental actors can use them. The data published on these portals are usually recorded in the form of metadata organized in rows and columns containing different information depending on the government sector bodies [72].

Generally, a dataset consists of one or more data files (known as resources) that can be accessed or downloaded in various formats. These resources may either be hosted directly on the associated data portal or provide links to external data sources [42].

Many governments worldwide have acknowledged the importance of OD data and

have set up national and regional open data portals to facilitate data sharing (Figure 2.2).

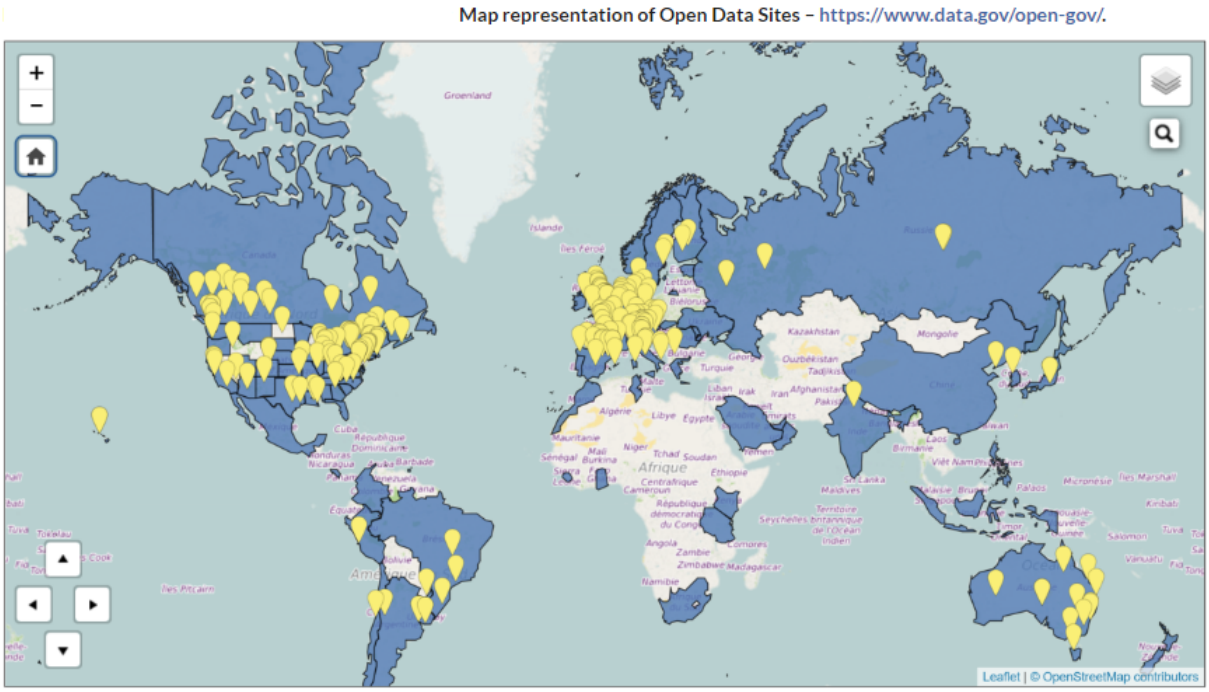


Figure 2.2. Global Open Data Portals
Source: Thompson [69, p. 112]

These platforms provide a wide array of datasets spanning different sectors such as healthcare, transportation, finance, and environment [8]. Some examples of national open data portal are depicted in the Table 2.14.

Table 2.14. National Open Data Portals (2023)

Country	Number of datasets (approx.)	URL
United States	250,000	www.data.gov
United Kingdom	50,000	data.gov.uk
Canada	80,000	open.canada.ca
Australia	70,000	data.gov.au
France	350,000	www.data.gouv.fr

Source : Campos Zabala [8, p. 308]

According to the European Commission[12], Open Data Portals (ODP) are online platforms created to simplify the search for reusable data. Similar to library catalogs, they store metadata entries of datasets intended for reuse, primarily focusing on raw numerical data instead of text documents. An example of a ODP is the Portuguese Open Data Portal, dados.gov.pt [13] (Figure 2.3). According to Mateus [36], there exists a significant proliferation of Open Data portals within the territory of Portugal. There are a total of twelve distinct open data portals (February 2023) that encompass various

[12]https://digital-strategy.ec.europa.eu/en/policies/open-data-portals
[13]https://dados.gov.pt/en/

domains. Several of these open data portals will be utilized as sources of data for our datasets pertaining to quality assessment.

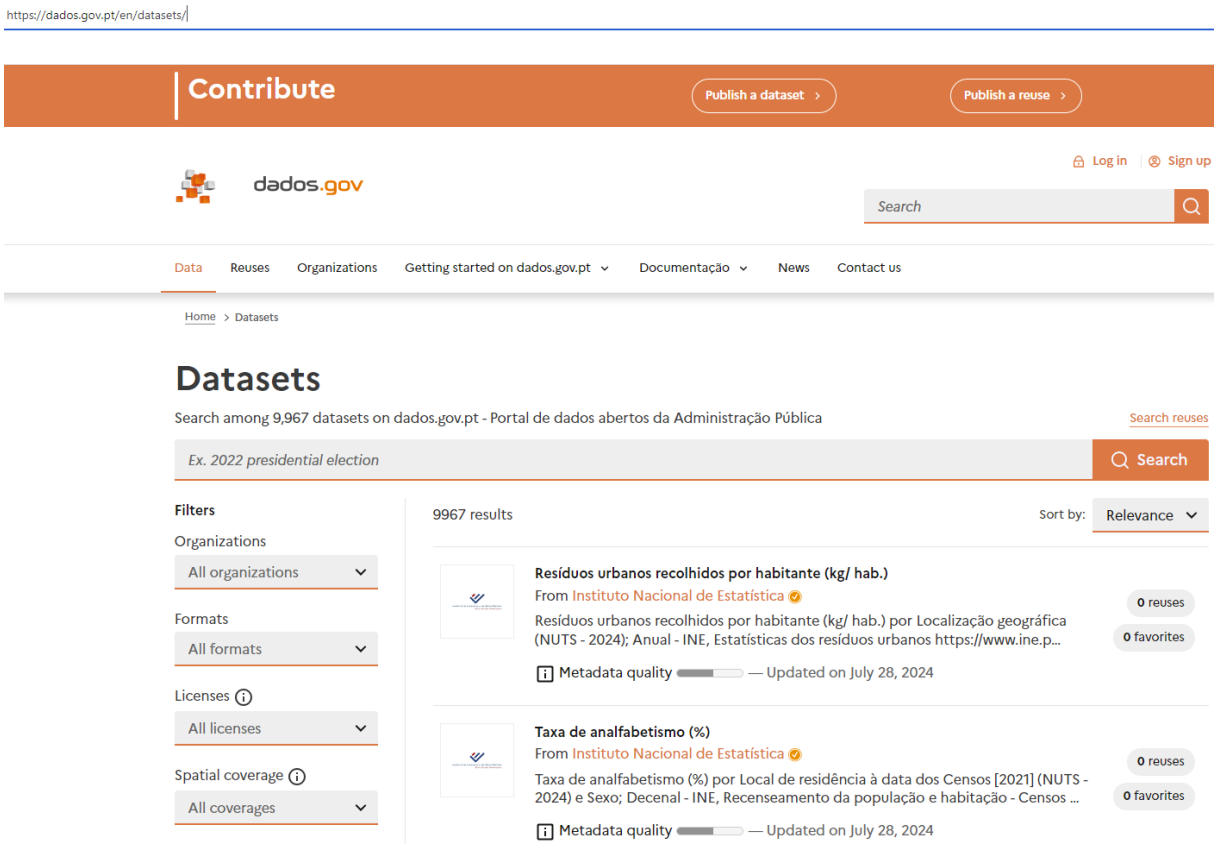


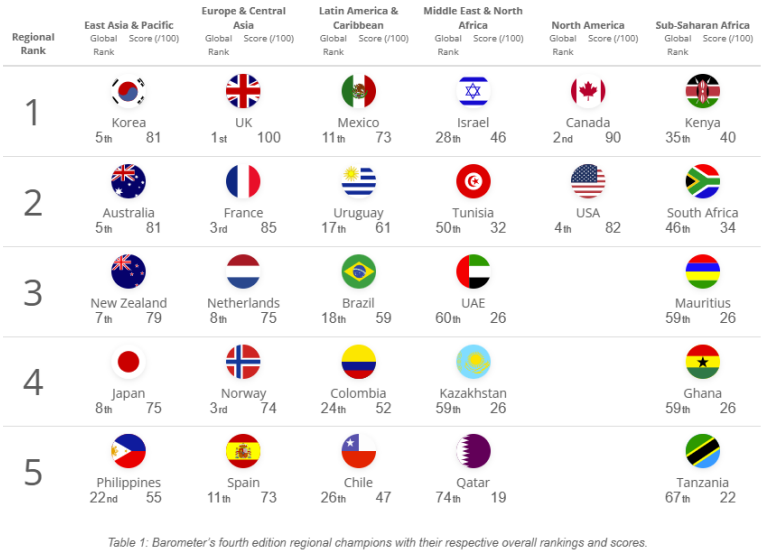
Figure 2.3. Portugal Open Data Portal
Source: https://dados.gov.pt/

The European Data Portal (EDP)[14], created and managed by the European Commission, collects metadata from open datasets that are accessible throughout Europe. Reviewing the datasets on the EDP portal is valuable for obtaining a comprehensive understanding of the quality of data being released by various European governments. Annually, the Open Data Maturity Report [19] provides insights into the level of maturity of OD in Europe, presenting statistical data on the content found on the EDP portal.

All members states of the European Union have a national portal in place to facilitate the easy discovery of openly accessible datasets [19].

The Open Data Barometer[15] monitors the progress of Open Data in 115 nations and territories, analyzing 15 different categories of government data such as maps, land, health, education, expenditures, companies and so on. According to the fourth edition of the Open Data Barometer Report, the leaders for each region in their study are Canada, Israel, Kenya, Korea, Mexico, and the UK (Figure 2.4).

[14]https://data.europa.eu/data/datasets
[15]https://opendatabarometer.org/



Regional Rank	East Asia & Pacific Global Rank	East Asia & Pacific Score (/100)	Europe & Central Asia Global Rank	Europe & Central Asia Score (/100)	Latin America & Caribbean Global Rank	Latin America & Caribbean Score (/100)	Middle East & North Africa Global Rank	Middle East & North Africa Score (/100)	North America Global Rank	North America Score (/100)	Sub-Saharan Africa Global Rank	Sub-Saharan Africa Score (/100)
1	Korea	81	UK	100	Mexico	73	Israel	46	Canada	90	Kenya	40
	5th		1st		11th		28th		2nd		35th	
2	Australia	81	France	85	Uruguay	61	Tunisia	32	USA	82	South Africa	34
	5th		3rd		17th		50th		4th		46th	
3	New Zealand	79	Netherlands	75	Brazil	59	UAE	26			Mauritius	26
	7th		8th		18th		60th				59th	
4	Japan	75	Norway	74	Colombia	52	Kazakhstan	26			Ghana	26
	8th		3rd		24th		59th				59th	
5	Philippines	55	Spain	73	Chile	47	Qatar	19			Tanzania	22
	22nd		11th		26th		74th				67th	

Table 1: Barometer's fourth edition regional champions with their respective overall rankings and scores.

Figure 2.4. The governments that are most engaged in the creation of Open Data
Source: https://opendatabarometer.org/4thedition/report/

Portugal's score is 42 of 100 in the Open Data Barometer Report 2016[16]. The overall score is a combination of readiness(58), implementation(47) and impact(16). Figure 2.5 depicts country details metrics.

2.3.3. Data Quality in the context of Open Data

According Carvalho et al. [10], it is essential to be able to evaluate the quality of open data if the goal is to reuse it. The potential of open government data can be negatively impacted by low data quality, making it difficult or even impossible to reuse [32]. End-users often assume that open data is of high quality [13]. The importance of the problem of OD quality can also be observed through the quantity of studies conducted. As per Scopus[17], the number of studies focusing on Open Data quality that were published in 2011 is 4.48 times lower than the number published in 2023. The results suggest that there has been a significant increase in the level of interest in Open Data quality since 2018, when both, the number of open datasets and the number of open data portals began to increase (Figure 2.6). According to W3C Data Catalogue Vocabulary (DCAT)[18] a dataset is defined as a *"collection of data, published or curated by a single agent, and available for access or download in one or more formats"*. According to European Union et al. [19], High-value datasets are datasets with a high potential economic and societal impact. EU Member States are currently in the process of determining key data areas that should receive priority for release datasets with a particular focus on statistics, geospatial information, earth observation, environmental data, and meteorological data.

[16]https://opendatabarometer.org/4thedition/?$_{y}ear$ = 2016*indicator* = *ODB*
[17]https://www.scopus.com
[18]https://www.w3.org/TR/vocab-dcat-1/

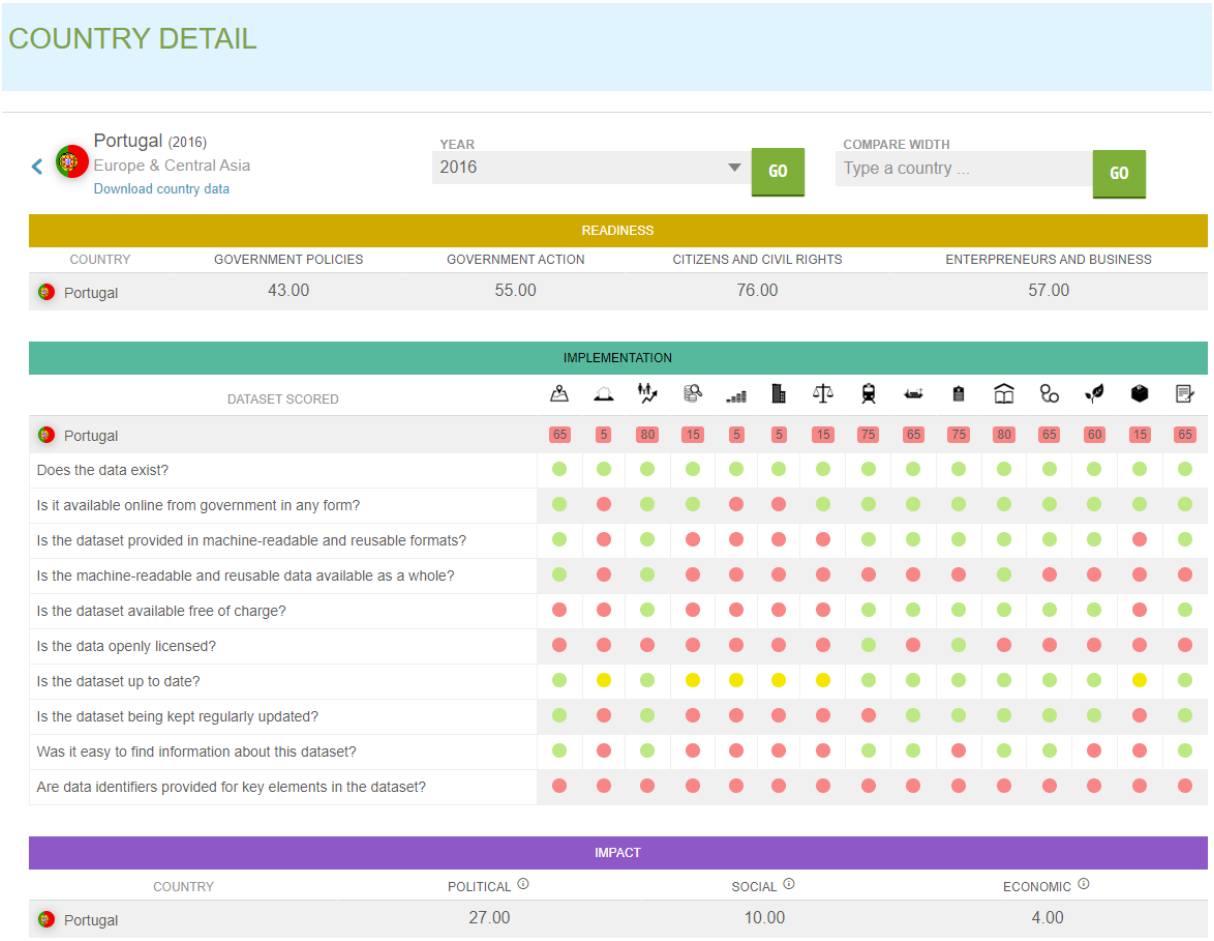


Figure 2.5. Open Data Barometer - Portugal
Source: https://opendatabarometer.org/4thedition/detail-country/?_year=2016&indicator=ODB&detail=PRT

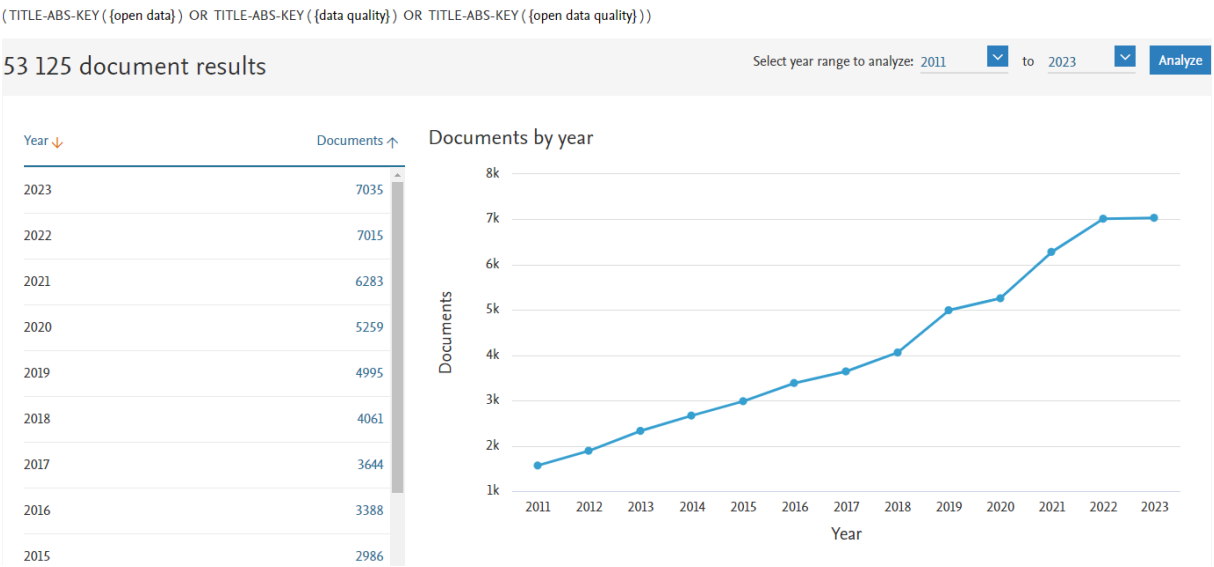


Figure 2.6. Open Data Quality studies by year of publication
Source: By the Author, data from Scopus, February 2024

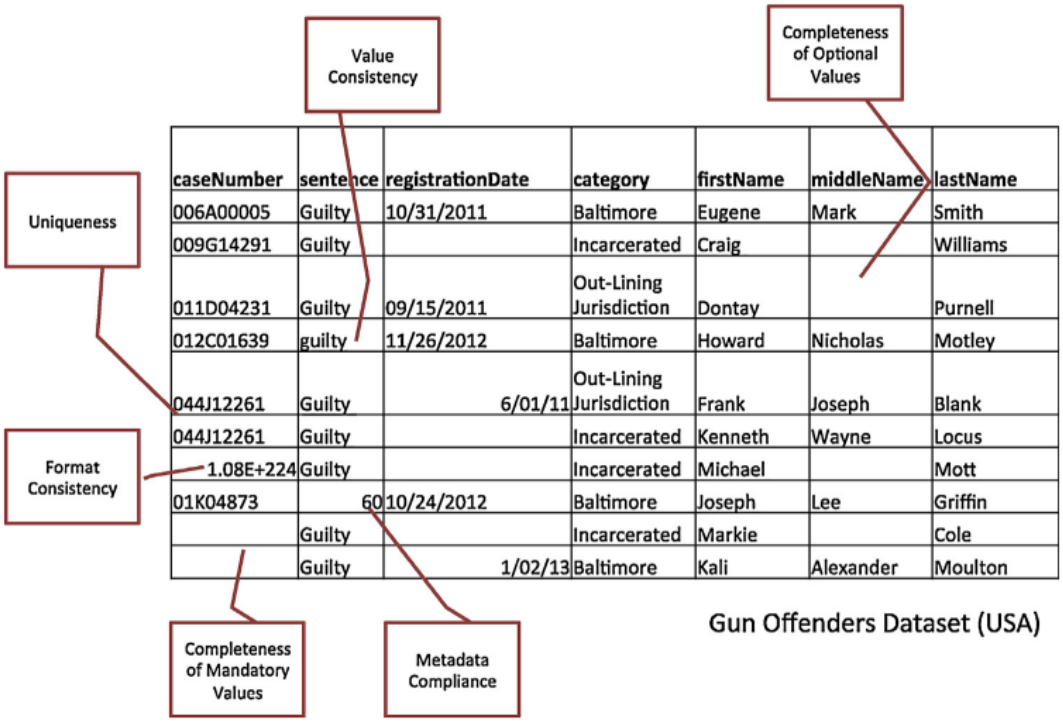
The Open Data (OD) movement has experienced a significant increase as governments have also implemented initiatives to disseminate data through OGDP and open datasets. For example, New Zealand, Canada, France, and the United Kingdom possess a significant array of datasets. The respective OGDP of the United States[19] offers more than 194,000 datasets. These statistics do not encompass the numerous organizational and other datasets provided by private sources such as GeoNames, Wikidata, and DBPedia, among others.

2.3.3.1. Aspects of Data Quality - Dimensions

Measuring and comparing the quality of OD is a complex task due to the need to consider multiple quality dimensions that may differ from each other, as well as the diverse range of stakeholders involved in OD [30]. One of the major concerns associated with the utilization of OD is the insufficient recognition of the inherent data quality. OD is frequently employed for a purpose that was not initially intended during the data collection process. Therefore, a dataset that may be adequate for one purpose may not be suitable for another purpose. Data consumers frequently encounter unfamiliar datasets that they may not have any ownership of [58]. Having access to OD datasets is crucial, but it becomes useless if we lack the capability to comprehend and analyze them [26]. Thus, the quality of OD may need to be approached differently compared to discussions about traditional data quality. Having access to OD datasets is crucial, but it becomes futile if we lack the capability to comprehend and analyze them [26].

Similar to other forms of information like structured data, OD encounters issues with quality such as inconsistency, inaccuracy, incompleteness, and completeness (Figure 2.7). Numerous published datasets experience quality problems like syntax errors, redundant instances, and incorrect or incomplete attribute values [5].

[19] https://data.gov/



caseNumber	sentence	registrationDate	category	firstName	middleName	lastName
006A00005	Guilty	10/31/2011	Baltimore	Eugene	Mark	Smith
009G14291	Guilty		Incarcerated	Craig		Williams
011D04231	Guilty	09/15/2011	Out-Lining Jurisdiction	Dontay		Purnell
012C01639	guilty	11/26/2012	Baltimore	Howard	Nicholas	Motley
044J12261	Guilty	6/01/11	Out-Lining Jurisdiction	Frank	Joseph	Blank
044J12261	Guilty		Incarcerated	Kenneth	Wayne	Locus
1.08E+224	Guilty		Incarcerated	Michael		Mott
01K04873	60	10/24/2012	Baltimore	Joseph	Lee	Griffin
	Guilty		Incarcerated	Markie		Cole
	Guilty	1/02/13	Baltimore	Kali	Alexander	Moulton

Figure 2.7. Data quality problems identified in an open dataset
Source: Sadiq and Indulska [58, p. 151]

Definitions of quality in the context of Open Data can vary significantly. Several different standards, methods, methodologies, and frameworks are employed for the purpose of assessing the quality of OD. The academic papers, identified a variety of dimensions used to measure the quality assessment of OD - It is important to highlight that a significant proportion of papers reviewed within this classification originates from the domains of Linked Open Data or open government data platforms. These dimensions encompass the quality of metadata, the accuracy of semantic and syntactic aspects, the uniqueness of the data, its level of completeness, consistency, accessibility, relevance, accuracy, reliability, clarity, timeliness and punctuality. Some authors concentrated solely on a single dimension, for instance Ali et al. [1], who examined the dimension of data completeness.

Various metrics are utilized to quantify the dimensions of data quality in the evaluation of OD. These metrics serve to assess the quality of the published data. The utilization of these metrics guarantees that the data disseminated through OD portals possesses a high level of quality, allowing users to effectively utilize it.

For instance, the European Data Portal (EDP) evaluates the quality of datasets concerning to the FAIR principles [2]. The FAIR principles are depicted in the Table 2.15.

The FAIR Data Principles (Findable, Accessible, Interoperable, and Reusable), which were introduced in Scientific Data[20] in 2016, represent a set of fundamental principles put forward by a group of scientists and organizations to promote the reusability of

[20]https://www.nature.com/articles/sdata201618

Table 2.15. FAIR principles

Principle	Description
(**F**)indable	Discoverable with metadata, identificable and locatable by means of a standard identification mechanism.
(**A**)ccesible	Always available and obtainable; even if the data is restricted, the metadata is open.
(**I**)nteroperable	Both syntactically parseable and semantically understandable, allowing data exchange and reuse between researchers, institutions, organizations or countries.
(**R**)eusable	Sufficiently described and shared with the least restrictive licenses, allowing the widest reuse possible and the least cumbersome integration with other data sources.

Source: https://www.go-fair.org/fair-principles/

digital assets. The principles refer to three categories of entities: data (or any digital object), metadata (information about that digital object), and infrastructure.

The FAIR framework offers a structure for categorizing the diverse elements of data quality. This framework comprises of four dimensions - Findability, Accessibility, Interoperability, and Reusability - and offers specific measurements for each dimension. The European Union [18] suggests 12 relevant indicators for data quality across the four FAIR dimensions (Figure 2.8).

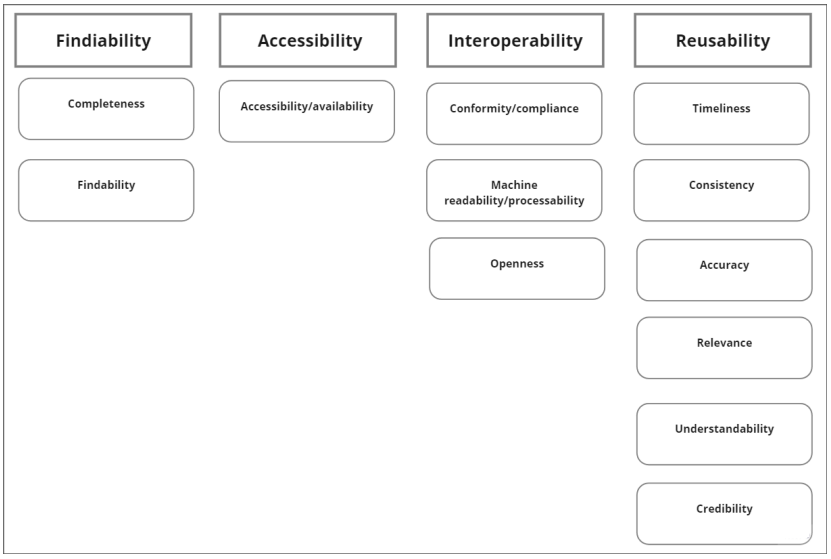


Figure 2.8. Overview of quality indicators grouped by FAIR dimensions
Source: European Union [18]

The ISO/IEC 25012 (previously described), is a comprehensive framework for assessing the quality of data. It encompasses various dimensions of quality, taking into account both the inherent nature of the data and its dependence on the system. When we talk about inherent data quality, we are referring to the extent to which data possesses the inherent capability to fulfill stated and implied needs. This is true when the data is used in specific conditions. On the other hand, system dependent data

quality pertains to the level at which data quality is achieved and maintained within a computer system.

Behkamal et al.[5] present a methodology based on ISO 25012 for the assessment of data quality pertaining to Linked Open Data (LOD) datasets prior to their publication, which encompasses the evaluation of quality dimensions or criteria that specifically concentrate on the intrinsic characteristics of data quality. Through a metrics-driven approach, presents inherent quality characteristics of LOD, which include semantic accuracy, syntactic accuracy, uniqueness, consistency and completeness and are displayed in Figure 2.9.

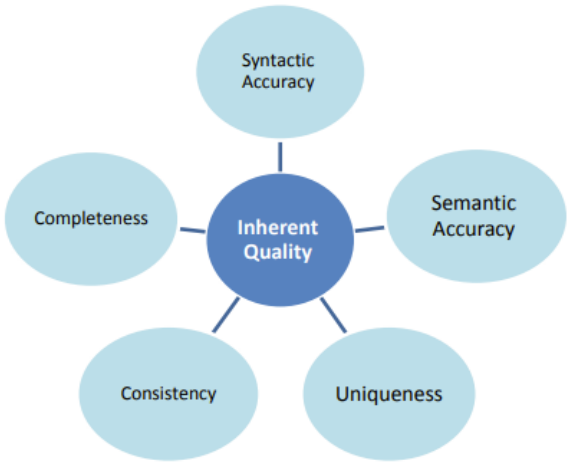


Figure 2.9. The proposed inherent quality characteristics of LOD.
Source : Behkamal et al. [5]

Furthermore, the metrics they utilize pertain specifically to linked data, which is structured in the form of triples, whereas our metrics are focusing to tabular data. LOD constitutes a small portion of the entirety of Open (Government) Data. More specifically, in the context of Portugal, the number of public datasets available in RDF format is 10, in contrast to the 956 datasets that are formatted in a tabular structure such as CSV, XLS, XLSX [54]. To operationalize the measurement of the five inherent quality characteristics mentioned above, a total of 20 metrics have been established. The process of defining metrics begins with the utilization of the Goal Question Metric (GQM) approach. It begins by outlining specific goals, formulating corresponding questions for each goal, and ends with an analysis conducted through theoretical validation and empirical evaluation.

Batini and Scannapieco [4] categorized dimensions into clusters (Accuracy Cluster, Completeness Cluster, Redundancy Cluster, Readability Cluster, Accessibility Cluster and Consistency Cluster) and defined dimensions for each one, e.g. Readability Cluster is associated with the Understandability dimension. After clustering dimensions, they look at how the dimensions related to one another.

Stróżyna et al. [68] developed a framework based on dimensions (Accessibility,

Relevance, Accuracy & Reliability, Clarity, Timeliness & Punctuality, Coherence & Comparability) for the quality-based selection of Open Data and evaluated an use case from the maritime domain.

Nikiforova et al. [46] suggest a method for assessing the quality of open data that includes examining various factors like the presence of values, alignment with data types, structure of stored values, adherence to specific patterns, alignment with enumerable values, and accuracy of values. The suggested solution for evaluating data quality comprises three key elements:

(1) **Data object definition.** Traditionally, the concept of a data object is commonly known as the collection of values of the variables that define an actual object in real life. For example, a simple illustration can be found in the Company Register of Latvia. In Figure 2.10, the data object "Enterprise" is represented along with its attributes: Reg_number – the registration number of the company, Name – the name of the company, Type – the category of the company, and so on.

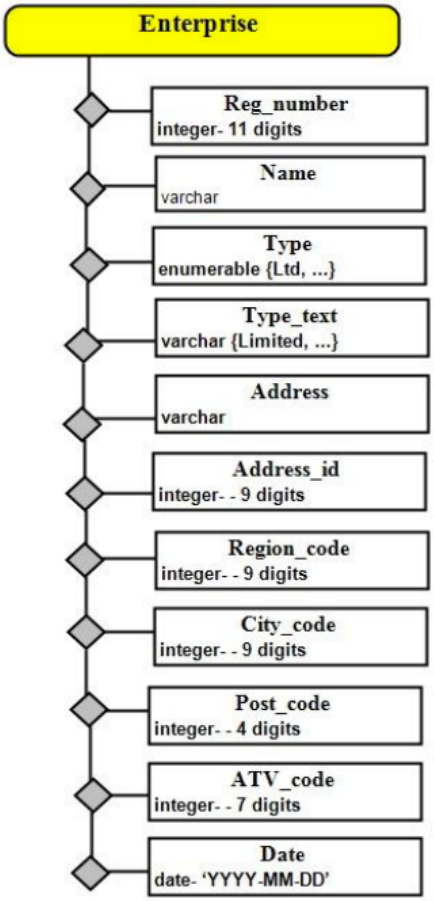


Figure 2.10. Data Object "Enterprise"
Source: Nikiforova et al. [46]

(2) **Quality requirements for data object.** A data quality specification includes requirements that need to be fulfilled for a data object to be classified as high

quality. The specification for quality (Figure 2.11) might consist of informal explanations of requirements, such as in plain language or formalized descriptions that are implementation-independent. The data quality specification for a data object is determined by logical expressions. The attributes/fields of a data object are used as operands in these logical expressions.

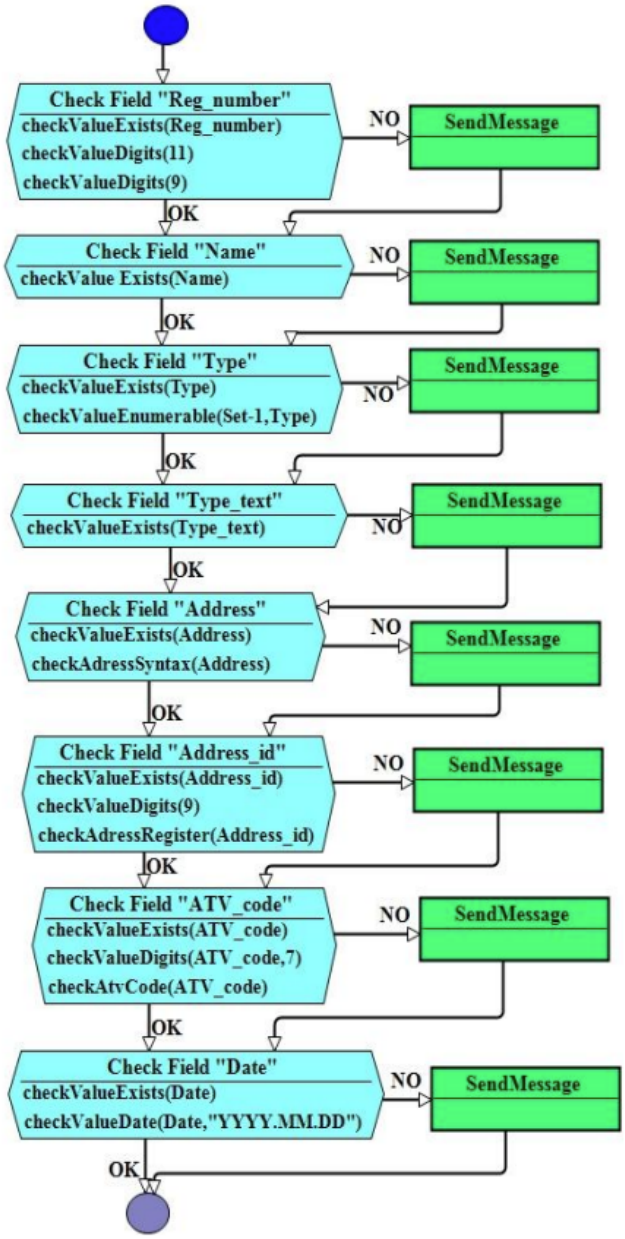


Figure 2.11. Quality Specification
Source: Nikiforova et al. [46]

(3) **The process of quality evaluation.** The initial phase of the quality assessment process outlines the steps to choose data object values from various data sources. Following that, additional steps are executed to assess the data object based on its quality, each focusing on a specific test for the conformity of the data object "Enterprise" with quality standards. The process of evaluating

quality involves verification tasks for each data object, which can be defined informally in natural language, through Unified Modeling Language (UML) activity diagrams, or in a DSL (Domain Specific Languages)[21]. Figure 2.12 displays distinct field checks for the Register data object, with each operation assessing field quality using a SQL statement. The SELECT statement in SQL identifies the target data object, while the WHERE clause specifies the quality criteria. This approach to data quality is commonly employed when data is stored in relational databases. The process of evaluating quality involves choosing values for data objects, assessing them against quality standards (Figure 2.11).

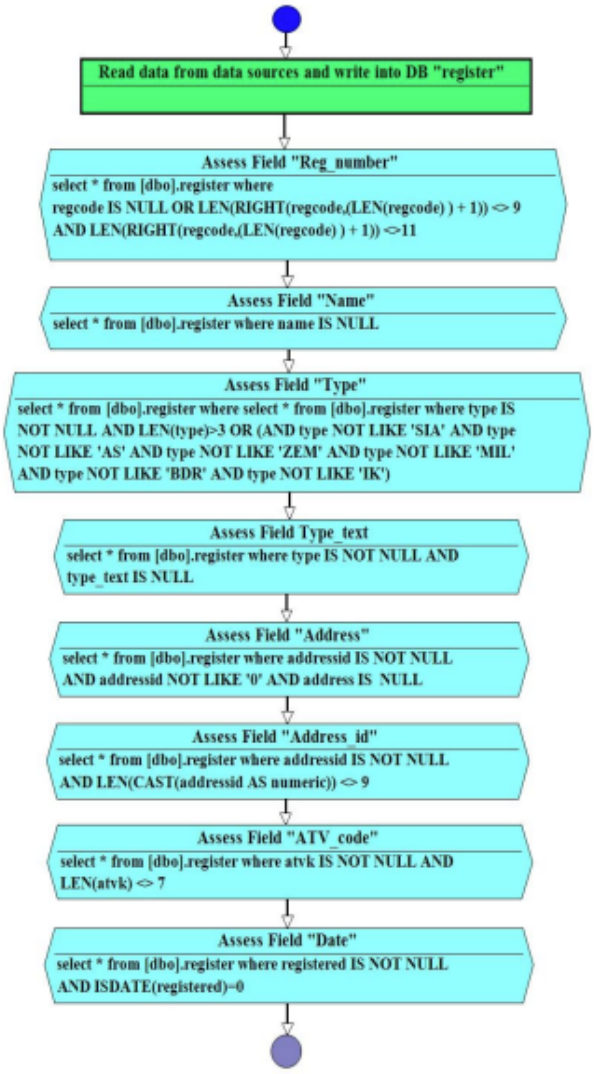


Figure 2.12. Quality Evaluation Process
Source: Nikiforova et al. [46]

[21]https://martinfowler.com/dsl.html

A fundamental principle of the approach is centered around data objects (preventing the connection of data quality problems by dimension). This examination is carried out on data releases from four European countries. This method requires a substantial amount of manual intervention and is dependent on the accurate and precise formulation of the data quality requirements.

Neumaier et al. [43], present a comprehensive overview of automated quality assessment frameworks, which facilitate the identification and evaluation of quality within open data portals. These authors propose five dimensions for evaluate Open Data Portal (ODP) : Existence (completeness), Conformance (Usage), Retrievability, Accuracy, and Open Data (Openess). The solution is helpful for assessing the overall quality of the portal, rather than open datasets.

The City of Toronto Open Data portal[22], evaluates the quality of data using five dimensions and give it a score for each dimension in order to weighing each dimension differently, to create a final score: Freshness (35%), Metadata (35%), Accessibility (15%), Completeness (10%) and Usability (5%), as outlined in Figure 2.13.

Quality Dimension	Weight	Metrics
Freshness Is this dataset up-to-date?	35%	- Has the data been refreshed on schedule? - Has the data been left unrefreshed for more than 2 years?
Metadata Is this data well described?	35%	- Are there metadata missing from the dataset? - Is the contact owner opendata@toronto.ca? - Is the "Learn More" URL a valid URL? - Are data definitions missing?
Accessibility Is this data easy to access for different kinds of users?	15%	- Are there any tags on the dataset? - Is the data updated manually or automatically? - Is the data stored as a file, or in the Open Data database?
Usability How easy is it to work with the data?	10%	- Do the columns have meaningful names? - Do columns have constant columns?
Completeness Is there lots of data missing?	5%	- Does the data consist of more than 50% null values?

Figure 2.13. Quality dimensions, weight and metrics.
Source: Open Data Toronto

After the final score is calculated, a grade is determined by utilizing established thresholds: 80% score and above gets Gold, 60% to 79% score receives Silver and everything else under 59% gets Bronze. The primary roles include assessing the level of data quality for each dataset before its publishing and furnishing the user with details regarding the data quality of the dataset.

In the realm of Artificial Intelligence (AI), Campos [8, p. 323] states *"One of the main concerns when using open data is the quality and reliability of the data. As open data comes from various sources, it is essential to ensure that the data is accurate, complete, and up to date to prevent the introduction of errors into AI models"*. It is crucial to verify that the data is precise, comprehensive, and current, as OD originates from diverse origins. This is

[22]https://open.toronto.ca/

necessary to prevent errors from being introduced into AI models. Indeed, OD can be of variable quality. Some open datasets are well curated and reliable, while others may be incomplete, inaccurate, or biased [8]. It is crucial to assess the quality of every open dataset before utilizing it. Campos [8] suggests to assess an open dataset, consider the following dimensions:

- **Accuracy**: Open data might include inaccuracies or mistakes caused by human error, outdated information, or misinterpretation of data. These inaccuracies have the potential to result in subpar model performance and inaccurate predictions.
- **Completeness**: Open data sources might contain missing or incomplete data, potentially harming the efficiency of AI models. The absence of complete data could result in biased models or overfitting.
- **Timeliness**: Open data may not always be up to date, which can affect the relevance of the data for AI projects. Outdated data can result in models that are not able to adapt to current trends or capture recent changes in the environment.

2.3.3.2. Data Quality Frameworks

The quality criteria of OD most commonly used so far are the ones outlined in the Five-Star Model [53]. Batini and Scannapieco [4] states that the quality of a dataset is evaluated by the five-star rating system based on its adherence to the principles of linked data.

To promote the publication of LD, Tim Berners-Lee, the creator of the LD approach, introduced a set of criteria to grade data quality on a scale from zero to five stars. This system allows data publishers to assess the extent to which their datasets adhere to the principles of linked data, as indicated by the rating system described in Table 2.16 .

Table 2.16. Tim Berners-Lee's Open Data 5 start rating

Level	Description	Benefits
★	Information is available on the Web (any format) under an **open license** to be *Open Data*.	Users have the ability to view, print, and save the data, in addition to manually choosing it on a system.
★★	Information is available as **machine-readable** structured data (e.g., Excel instead of an image scan of a table).	Data can be automatically processed, visualized, and converted into different formats.
★★★	**Non-proprietary** formats are used (e.g., Comma-separated values (CSV) instead of MS Excel).	Data can be altered regardless of its format and without being dependent on any specific software.
★★★★	URI identification is used so that people can point at individual data.Use **open standards** (RDF,SPARQL) to identify things.	Data has the ability to be connected, labeled, and utilized again.
★★★★★	**Data is linked** to other data to provide context.	Data patterns can be automatically identified and it is possible to dynamically uncover complementary data related to the original data.

Source: Adapted from https://www.w3.org/DesignIssues/LinkedData.html

The 5 Star Open Data rating system[23] is considered as a way to identify high-quality data, focuses primarily on Linked Data (LD). The model establishes five levels of quality for OD (five stars) with a focus on two key aspects: Availability (levels 1, 2, and 3)

[23]http://5stardata.info/.

and Reusability (levels 4 and 5). The schema proposed by Tim Berners-Lee's approach covers only specific aspects, e.g. format or encoding used to publish the data and do not cover other aspects related with data values itself. This means that a dataset can be graded with 5 starts while data values have poor quality [73].

On the other hand, the G8 Open Data Charter[24] and the Open Data Institute Certification Badges[25] primarily focus on aspects such as use of a robust and consistent metadata, data schema descriptions, discovery, licensing, instead of the data [22].

The Center for Open Data Enterprise[26] has recognized the importance of its data for a long time, and they have developed the Federal Data Quality Framework to give priority to its quality, considering a few common elements (dimensions) that can be used to describe the data quality: Accuracy, Metadata, Machine-readability, Timeliness, Granularity and Interoperability.

Vetrò et al. [73] present a quality framework based on the SQuaRE (ISO/IEC 25012, 2008) standard, as well as in their literature review. Data quality dimensions and metrics are use for assessment, distinguishing itself by focusing on Open Government Data. The framework put forward by Vetrò et al. [73] include accuracy, accessibility, completeness, timeliness, consistency, and understandability.

Krasikov and Legner [28] propose a methodology based on Design Science Research (DSR) to evaluate, analyze, and organize open data for application within a corporate environment. This approach for assessment OD focuses on metadata quality and dataset content. Completeness, Uniqueness, Validity (format compliance) are key dimensions in open data assessment.

Zhang et al. [75] proposed a methodology called LANG, which is established using a Design Science approach based on semiotics theory and data quality dimensions. The LANG approach assists in identifying data quality issues by assisting data users in exploring datasets with unfamiliar sources. The process is divided into two main stages: namely, the syntactic stage which centers on data quality problems related to syntactic constraints in the data (Uniqueness, Format Consistency, Referential Integrity, Meta-data Compliance, Business Rule Compliance); and the semantic stage, which concentrates on identifying data quality issues linked to data semantics (Completeness [Mandatory Attributes], Completeness [Optional Attributes], Semantic Consistency, Value Consistency, Precision, Non-Redundancy).

These approaches aim to tackle the various challenges associated with identifying appropriate open datasets and making them ready for utilization. Existing approaches to data quality management are inherently hierarchical. In this approach, data quality (DQ) needs are identified from a top-down perspective, based on well-established usage requirements, and are then enforced through effective data governance practices

[24]https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex.
[25]https://certificates.theodi.org/en/about/badgelevels.
[26]http://reports.opendataenterprise.org/BriefingPaperonOpenDataandImprovingDataQuality.pdf

[58]. However, these methods usually rely on datasets that are within organizations, containing metadata that is known and an understanding of the data's meaning. On the contrary, OD often unfamiliar to the user and may not have metadata [58]. An outline of the quality dimensions is provided in the table table 2.17.

Table 2.17. Overview of Quality Dimensions for Open Data in principal studies

	Principal Studies References							Organizations		N°
Quality Dimension	**Batini and Scannapieco [4]**	**Behkama et al. [5]**	**Neumaie et al. [43]**	**Stróżyna et al. [68]**	**Vetrò et al. [73]**	**Zhang et al. [75]**	**Krasikov and Legner [28]**	**European Union - FAIR principles [11]**	**Toronto Open Data [36]**	**Studies**
Completeness	✓	✓	✓		✓	✓	✓	✓	✓	8
Accuracy	✓	✓	✓	✓	✓			✓		6
Consistency	✓	✓				✓		✓		4
Understandability	✓		✓		✓			✓		4
Accessibility	✓			✓				✓	✓	4
Timeliness	✓			✓	✓			✓		4
Uniqueness		✓				✓	✓			3
Usage			✓						✓	2
Interoperability	✓							✓		2
Openness			✓					✓		2
Conformance					✓		✓	✓		2
Relevance				✓				✓		2
Metadata						✓			✓	2
Comparability				✓						1
Reusability								✓		1
Traceability						✓				1
Findability								✓		1
Retrievability			✓							1
Freshness									✓	1
Credibility								✓		1
Processability								✓		1

2.3.3.3. Tools

Data quality tools basically establish the domains of permitted data values and impose restrictions, usually using business rules [15]. According to Gartner[27], data quality tools are "*the processes and technologies for identifying, understanding and correcting flaws in data that support effective information governance across operational business processes and decision making*". Gartner has examined 16 tools provided by popular vendors and categorized them within a Magic Quadrant for Data Quality Solutions (formerly Magic Quadrant for Data Quality Tools), distinguishing between 'Leaders', 'Challengers', 'Niche Players', and 'Visionaries' as depicted in the Figure 2.14.

[27]https://www.gartner.com/en/information-technology/glossary/data-quality-tools

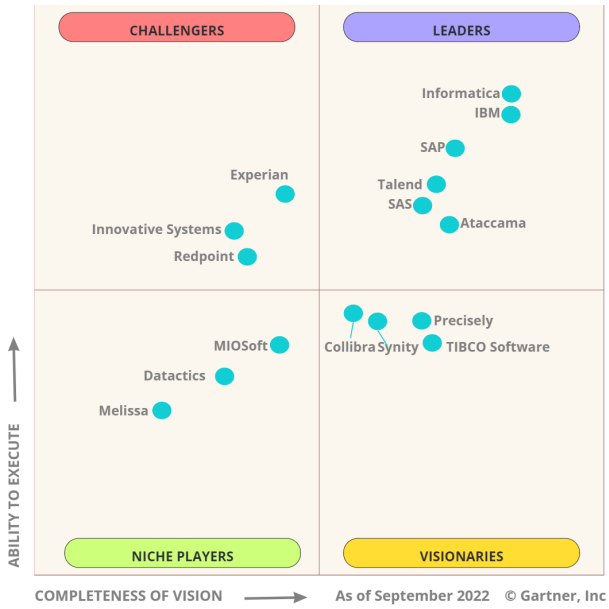


Figure 2.14. Magic Quadrant for Data Quality Solutions.
Source: Gartner (November 2022)

This quadrant is an overview in time of the vendor's tools available in the market and their ranking based on Gartner's criteria[28]. The four quadrants and their respective short meanings are as follows:

- **Leader**: supply established products that satisfy consumer needs and have proven to have the vision required to maintain their market position when needs change.
- **Challenger**: demonstrates execution skills but lacks a clear plan for the future and market understanding.
- **Visionary**: possesses a clear understanding of the market but has not demonstrated its ability to execute.
- **Niche Player**: concentrates on a specific segment of the market but lacks a defined execution plan and strategy.

Data profiling and exploration are the main focus of these commercial quality tools. The wide variety of tasks that these tools focus on include data redundancy checks, column correlation analysis, validity checks, functional dependence analysis, and statistical distribution analysis of data. The drawbacks of these tools are data dependency constraints, specialized towards specific dimensions and assumption relating to the availability of metadata [58].

Probably the most popular tool for data quality is Microsoft Data Quality Services (DQS) [46]. DQS is a data quality tool driven by knowledge, developed as a SQL Server component for the purpose of analyzing and enhancing data quality. According to Microsoft [29], DQS enables to build a knowledge base and use it to perform a variety

[28]https://www.gartner.com/en/documents/3956304
[29]https://learn.microsoft.com/en-us/sql/data-quality-services/

of critical data quality tasks, including correction, enrichment, standardization, and de-duplication of data, data matching, data cleansing and data profiling. There are essentially two main steps involved in working with Data Quality Services as outlined in Figure 2.15.

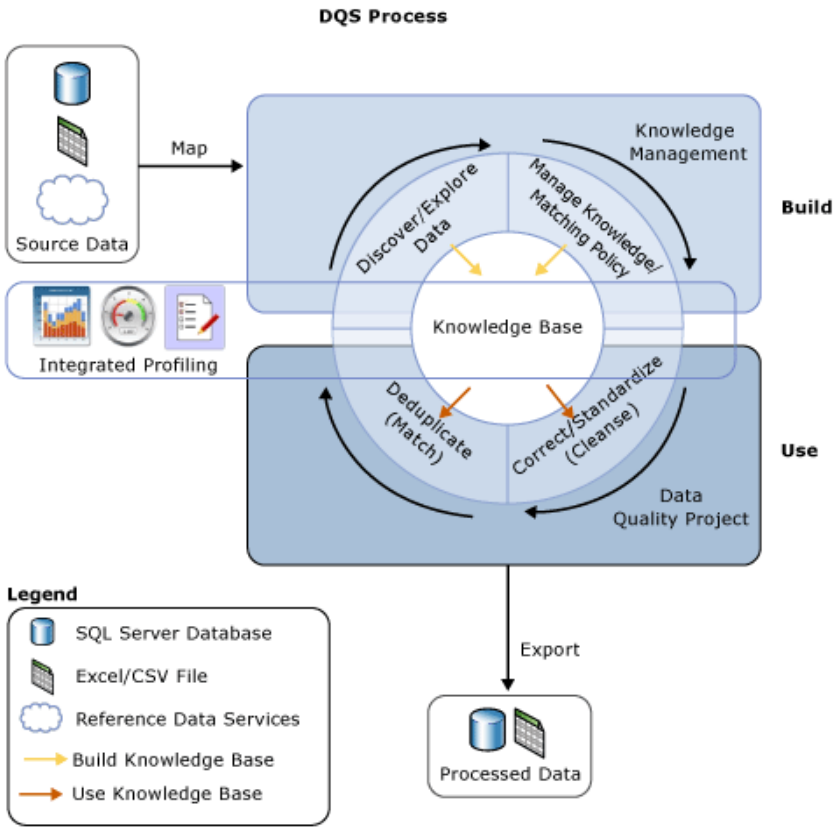


Figure 2.15. Data Quality Service (DQS) process.
Source: Microsoft [38]

Knowledge Management. First, you must create one or more knowledge bases (KBs). Within a KB, you store all information pertaining to a specific set of data - such as customer data. Each KB contains: 1) Domains that define valid values and correction rules for data fields, 2) Matching policies that define rules for identifying duplicate data entities. This information from the KB is utilized by DQS to detect inaccurate, incomplete, and invalid data, thereby enhancing the overall integrity and quality of the data. The knowledge base can be established through automated means (using knowledge discovery from sample data) or through manual input, allowing for flexibility and scalability.

Data Quality Project. Once you have finished building a knowledge base, you can create a DQS project indicating what KB will be use and the data source containing the data to be cleansed and map the columns in it to the domains in the KB. In DQS projects, you perform cleansing, profiling, and matching activities. After analysis, you will obtain a list of all correct values, incorrect values, suggested values and new values.

DQS presents some disadvantages, namely: it requires high resources, is pricey for small businesses, it is possible to analyse only one table per time (multiple table analysis is not available); has recently been integrated with the Microsoft Azure ecosystem but Microsoft has not announced any significant updates or improvements in recent years.

Great Expectations[30], is an open-source Python library created to improve data quality and testing. The tool offers a strong foundation for validating, documenting, and profiling data, playing a crucial role in upholding high data quality standards. The following diagram (Figure 2.16) illustrates the end-to-end Great Expectations data validation workflow.

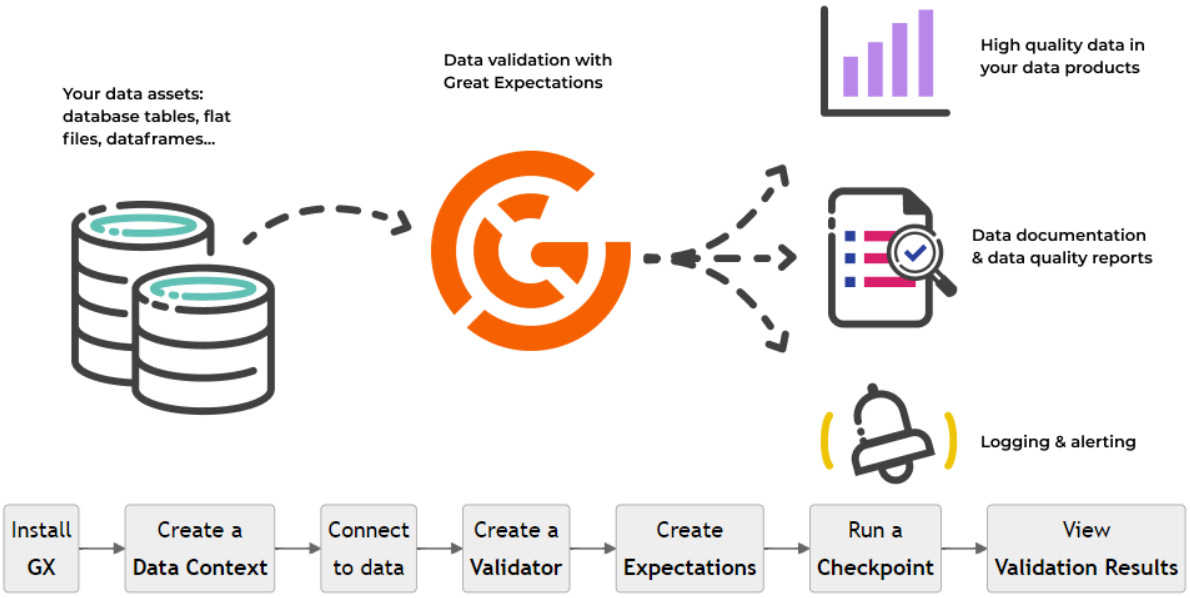


Figure 2.16. Great Expectations - Data validation workflow.
Source: https://docs.greatexpectations.io/docs/oss/tutorials/quickstart

This tool is accessible to everyone without any financial cost, easy integration with different data sources since flat files to dataframes and usefull data quality reports are generated at the end of the validation process. However, users with limited technical IT skills may face challenging when using Great Expectations.

According to Great Expectations, expectations[31] are assertions about the data, expressed in a declarative language. Expectations serve as a means of communication for discussing the attributes of data and its quality - facilitating interactions among humans, between humans and machines, and among machines. Consequently, expectations serve as indicators for assessing the quality of the data.

Their application in open data quality context to evaluate the quality of datasets according dimensions and metrics can tell us how feasible is the evaluation tool for open data quality assessment. Data quality tools essentially implement controls, typically

[30]https://greatexpectations.io/
[31]https://greatexpectations.io/expectations/

using business rules, that define the domains of allowed data values.

In conclusion, studies have determined that open data faces challenges with data quality. There exist tools, frameworks and academic studies that can be utilized for the evaluation and enhancement of data quality; however, they come with their own set of constraints and drawbacks. Generally, all of these works primarily focus on data quality problems before the data is published, and none of them provides a solution for identifying the quality problems in published datasets. It is important to note that most of the papers reviewed reveal a clear lack of a comprehensive theoretical framework, with most assessments focusing on open government data portals (OGDP) or Linked Open Data (LOD) rather than individual open datasets. It is crucial to investigate this further and propose a framework that could enhance it.

CHAPTER 3

Proposing of Framework

As described in the previous section (Table 2.17), the principal studies identified 21 quality dimensions used for assess OD. This section examines these different dimensions in order to propose a framework for the measurement of OD Quality that focuses on structured datasets in terms of data quality dimensions and its metrics. A dimension of data quality includes one or more metrics as depicted in Figure 3.1.

Most of the studies analyzed [4], [5],[28],[43], [73],[75], [19], [70] show that Completeness is one of the main dimensions used to assess the quality of open data. According to ISO [23], data itself has inherent quality independent of its context, so this dimension indicates the absence or presence of data. The data can be either mandatory or optional, a feature that is defined by the business rules, but the data user can infer whether or not the field should be with null values or with some data that implies absence of data, e.g., N/A or Unknown. Ali et al. [1], who assessment 20 open datasets, state that missing values are a critical issue in decision-making systems and can skew results.

Another frequently used metric [5],[75],[28] is the Uniqueness, Behkamal et al. [5] evaluate duplicate entries within datasets, which may distort analysis and insights. Zhang et al. [75] define Uniqueness of an attribute can be determines by the presence of duplicates in the data. One way to guarantee uniqueness in a relational database is to enforce constraints on primary keys [75]. Given that numerous datasets accessible from the open data portals comprised a singular table (which can be downloaded in formats such as CSV, XLSX, or JSON), the dimension of Uniqueness, predicated on key fields, was considered within this framework

In the literature [4], [68], [19], [70], [66] Accessibility has received significant attention as a crucial quality characteristic. Sebastian-Coleman [62] states that some dimensions of data quality depend on information that is part of the data ecosystem (e.g. reliability of systems, metadata) rather than available through the data itself. The possibilities offered by technology should be utilized to assess Open Data [65]. Data accessibility is particularly significant when it comes to open data, it permits the possibility to retrieve data from a source; it covers elements like the source's structure, the technologies employed, the source's stability (errors, unavailability of a service); it also considers login or registration requirements, as well as data unavailability at the time when needed [57]. The ability to effectively search, retrieve, and use data at all times can be considered a key characteristic of good data quality [67]. Song et al. [66] define Accessibility as a subordinate attribute of Availability. According to Monino and Sedkaoui [40], to develop knowledge, OD must be freely accessible online.

Furthermore, Interoperability is an important quality characteristic which has received attention in the literature [4], [19], [43] and it is also an important quality aspect of Open Data. Neumaier et al. [43] consider in the Openness dimension characteristics as Open Format, Machine-Readable and Open License. Krasikov and Legner [28] observed that the same quality characteristics are typically defined differently, therefore, both dimensions are related. Interoperability of open datasets is a key issue problem because the idea of open data is that it may be reused and connected to other data. The diversity of datasets, for example, in terms of data formats employed, therefore greatly increases the processing effort required for additional data consumption, and it can even make data reusability impossible [31]. Another aspect to permit the reusability, therefore, interoperability, is the Licensing. In the era of open data, licensing is a new quality component that is required [4]. Providing a transparent and accessible license is essential to facilitate the reutilization of a dataset. The licensing information may be presented in textual form on the official open data portal and as machine-readable metadata within the dataset itself [9].

The other quality dimensions mentioned in the Table 2.17 are not included in our proposed framework. For example Credibility describes the extent to which "data has attributes that are regarded as true and believable by users" ISO [23], therefore, the user's impression has a significant influence on this dimension. Different names for dimensions are utilized to represent the same meaning [48]. Freshness and Timelines is related with the frequency of updating metadata or data. The evaluation of the data's timeliness is not simple since it can be challenging to infer from the content whether the data is historical or current [19]. For determining Usage and Relevance necessitates knowing the usage context [58].

Based on a review of the methods for measuring data quality that have been published in the literature, an assessment framework is suggested in this context. The suggested framework will be focused on data consumers rather than data producers. According to Zhang et al. [75], data consumers have no influence over the creation and management of the data, it means, they are disconnected from data producers. The data producer has generally generated the data for the exclusive utilization of the organization, and there is not inherently a necessity to take into account the requirements of open data users by default [25]. Another important point is the absence of domain knowledge by data consumers. Hence, the framework proposed is oriented to the assessment of published OD instead of OD before publication focusing in structured dataset or any form of data than can be presented in a structured format.

The proposed framework uses the dimensions that address the fundamental inherent characteristics of data quality (Uniqueness, Completeness) supported by automatic computation of its metric(s) using IT Tools (Python and Great Expectations). A set of quantitative indicators are also proposed defined at different levels of detail: at the most granular level of measurement, that is row/cell, for the dimensions of Uniqueness

and Completeness, and the dataset level for the dimensions of Accessibility and Interoperability. Additionally, the suggested dimensions of the framework are categorized into two distinct groups in accordance with the quality characteristics delineated previously. The initial group (Accessibility, Interoperability) is predicated upon the systemic and technical facets of the open data environment, while the subsequent group (Completeness, Uniqueness) encompasses two inherent-related attributes of an open dataset. Intrinsic dimension denotes the quality of data as independent of the user's context. A set of common basic data quality metrics that can reveal details about specific dataset indicators is crucial for the intrinsic quality of the data [20].

An overview of the proposed data quality assessment framework is depicted in the Figure 3.1.

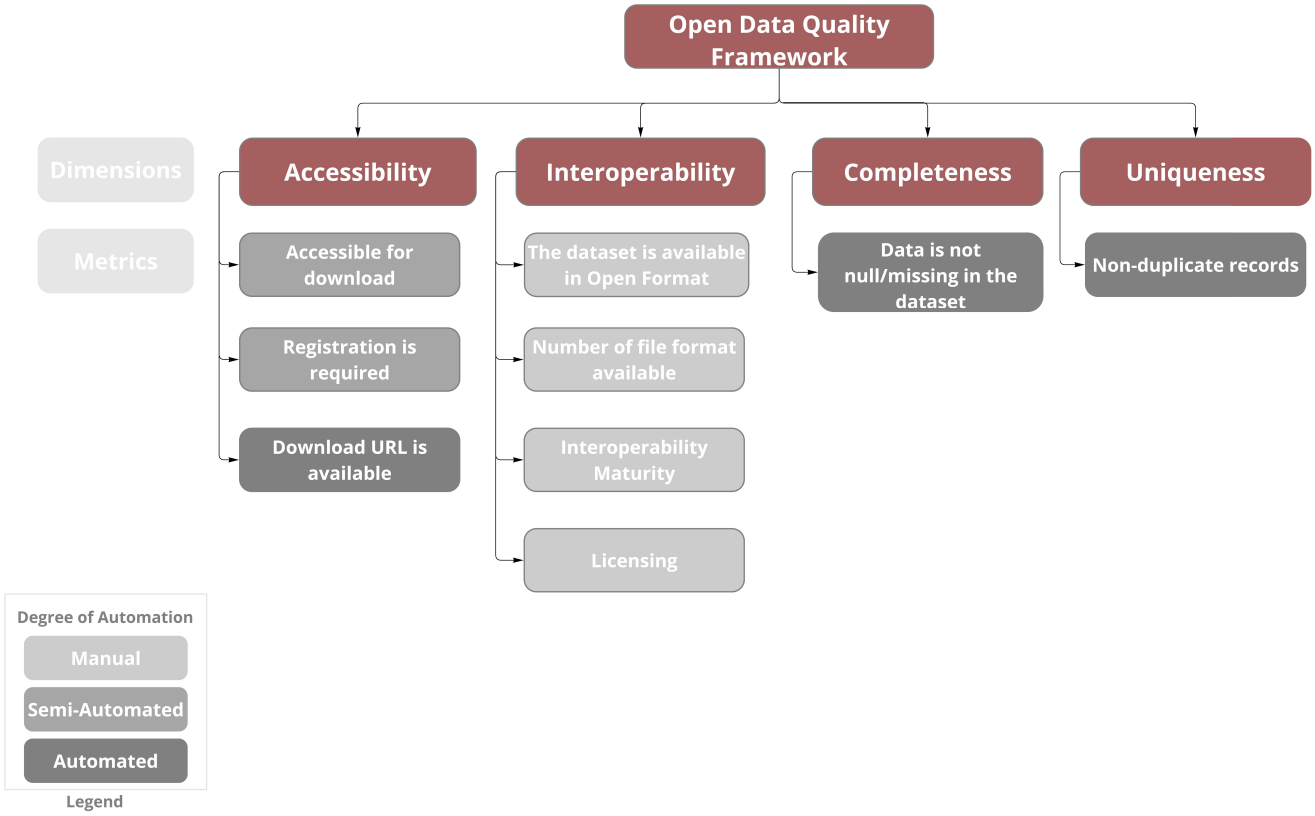


Figure 3.1. Open Data Quality Framework proposed
Source : By the Author

3.1. Quality Scoring

The Quality Score serves as a crucial element of the proposed final Framework, which consolidates all calculations related to the quality dimensions outlined within the framework (Figure 3.1) and provides an assessment of data quality by utilizing the suggested metrics, grounded in the parameterization of the dimensions.

The overall quality score is calculated using a range of 1 to 100 points. Each of the four dimensions has a potential high score depending on their relative importance as

shown in Table 3.3.

Table 3.1. Score per Dimension

Dimension	Points
Accessibility	30
Interoperability	30
Completeness	25
Uniqueness	15

Metrics are used to quantify the level of quality of each dimension. Depending on the quality dimension to be assessed and the chosen quality metric, a score to be evaluated by the user in the process of deciding on the suitability of the data for some intended use according the values depicted in Table 3.2. The score for each metric will be calculate in an Manual/ Semi-Automated/Automated way. For the automated way, an IT tool developed in Python and Great Expectations library was developed in the context of this dissertation (see section 3.3) will be use.

Table 3.2. Proposed scoring for the dimensions and metrics of the framework

Dimension	Description	Points	Metrics	Score (To-Be)
Accessibility	Dataset can be accessed by a human or computer without encountering any errors or access restrictions	30	* Accessible for download *Registration is necessary *Download URL is accessible	60 15 25
Interoperability	The extent to which data can combine with other data and work together	30	*The dataset is available in Open Format *Number of formats available *Interoperability Maturity *Licensing	70 10 10 10
Completeness	The data is considered complete when it contains all necessary items for representing the entity	25	*Data is not null/missing in the dataset	100
Uniqueness	A dataset should contain only one record for each instance of an entity	15	*Non-duplicates records	100

Source: By the author

3.2. Proposed Dimensions and Metrics

Each data quality dimension captures a specific quantifiable aspect of data quality [35]. The data quality dimensions that are more frequently utilized have been examined in the preceding literature review. The proposed Framework includes multiple assessment dimensions such as accessibility, interoperability, completeness, and uniqueness. Due to the fact that most data consumers deal with datasets that have little to no metadata about the schema, business rules, standards, or other relevant information, the dimensions that have been selected primarily center on dataset availability and data values quality in terms of completeness and uniqueness.

Evaluating the quality of open data involves the formulation of suitable metrics for assessing data quality [64]. Each dimension is subdivided into smaller components. These subdivided components constitute the second tier of data quality evaluation

known as Metrics. Other names that are used in relevant studies to describe this level are sub-dimension, dimension attributes, and criteria. Metrics are aggregated to dimension scores, which are then weighted individually towards the total quality score.

Because dimensions would contribute to the total quality score, the weights had to be assigned. Weights were determined based on aligned with the core principles of Open Data, access for reuse under licensing, Accessibility and Interoperability are the dimensions that permit assessment if the dataset complies. Uniqueness and Completeness, oriented to assess the quality of data values, can be determined by the presence of duplicates or missing or null in the data, in the opposite, other dimensions as Usefulness and relevance can only be determined when the use of context known.

3.2.1. Accessibility dimension

Refers to whether the dataset can be accessed and obtainable by a human or computer without encountering any errors or access restrictions. Data consumers can download the dataset via the open portal in ways other than direct download. If the data is inaccessible, the data consumer is unable to evaluate other dimensions of data quality such as completeness, uniqueness, and more [35].

The ability to search, access, and use data efficiently and consistently is a key characteristic of high-quality data [67]. Accessibility refers to whether the content of the portal or the resources can be accessed by a human or computer without encountering any errors or access restrictions [18].

Batini and Scannapieco [4, p.106-107] consider Availability as an element of the Accessibility dimension. Dataset downloads should not be prevented by technical barriers [43].

One fundamental principle of OD is its accessibility: data must be easily accessible and made available to a wide range of users to prevent restricting its potential for reuse. Discoverability and accessibility of Open Data are essential for its adoption and utilization [74]. Three metrics for Accessibility dimension assessment were defined as depicted in Table 3.3. In order to evaluate how simple it is for users or computer to obtain the datasets that are made available on Open Data portals, Accessible for download metric was defined with a score from 0 to 60 points (maximum score) according the criteria defined in Table 3.4. To assess that users can download without registering, we proposed the metric Registration is necessary scoring 15 points. This metric is penalized with 0 points if previous register is needed. The registration activity can delay the dataset's consuming but not limit the reuse. Usually, the primary point of access is a download URL, which needs to be published in the OGDP and accessible, meaning it can be accessed through a browser. The metric Download URL is accessible permits assess if there is any technical issue as such as temporally system downtime, networking and so on as depicted in Figure 3.2. This metric is penalized with 0 points URL is not available, otherwise 25 points are scored.

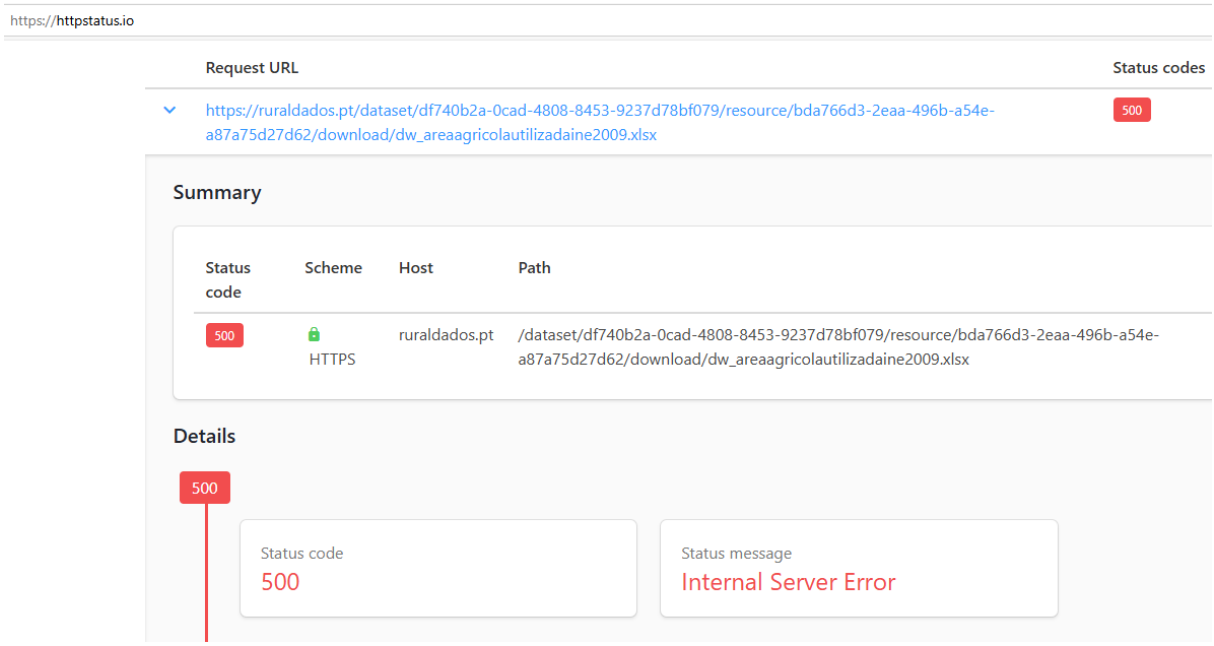


Figure 3.2. Download URL is not accessible

Table 3.3. Proposed metrics for Accessibility assessment

Metric	Description	Degree of Automation	Score (To-Be)
Accessible for download	Dataset can be freely downloaded	Semi-Automated	60
Registration is necessary	Registration is necessary in order to access the dataset	Semi-Automated	15
Download URL is accessible	The URL should be working and accessible	Automated	25

Source: By the author

3.2.1.1. Metric : Accessible for download

This metric defines the ease of downloading the dataset without complex procedures. According to European Union [18], to ensure convenient access and further processing, no limitations should be imposed, e.g. slow download or unresponsive. Beno et al. [6] point out that a large number of data consumers choose to use an application programming interface (API) because it allows for the direct integration of data into their applications, removing the need for manual downloading. The API serves as a software intermediary facilitating communication between two applications. The criteria for scoring this metric is depicted in Table 3.4.

Table 3.4. Accessible for download score

Score	Criteria	Comment
0	No download file	
30	Manual Download - individual file	Aggregated data can generate large amount of data and time-consuming download
40	Manual Download - set of files	Separate dataset can help split data and users can download files they need (e.g. Year)
50	API-download dataset automatically	Dataset can be consumed through application programming interfaces (APIs)
60	Manual Download and via API	Dataset can be download manually or consumed through API

3.2.1.2. Metric : Registration is necessary

This metric evaluate if prior registration is necessary to access the dataset and how it affects ease of access. According to European Union [18], in order to facilitate simple utilization, it is essential that there are no limitations to access, e.g. registration in the open data portal. According Beno et al. [6], certain portals restrict access to data unless the individual completes the registration process on their website. Criteria and scoring are depicted in Table 3.5.

Table 3.5. Registration is necessary score

Score	Criteria	Comment
0	Registration is needed for download dataset	Mandatory register limit accessibility to the dataset
15	Accessible for download without previous registration	A dataset is obtainable directly by clicking 'download' or via API

3.2.1.3. Metric : Download URL is accessible

This metric is used to determine the extent to which the dataset can be downloaded using the specified URL. Dataset's accessibility in relation to its up-time is critical. For instance, when an HTTP-GET request is send via browser for download a dataset, the status code "404 Not Found" is not returned (Batini and Scannapieco [4, p. 107]). Criteria and scoring are depicted in Table 3.6.

Table 3.6. Download URL is accessible score

Score	Criteria	Comment
0	URL given return error	HTTP status code between 400 to 500 indicate errors.
25	Dataset can be downloaded using URL given	The URL for downloading the dataset is working and accessible.

3.2.2. Interoperability dimension

Interoperability enables exchange of data and allows datasets to be reused in various application fields, focuses in types of file formats, metadata descriptions of the datasets and licensing. To ensure that the definitions and meanings of the data elements are clear, some information about schema or data fields should be included with the data. Metadata facilitates data cross-reference and interoperability, therefore, enhancing the value of information for reuse [22].

The dataset should be compatible with different systems and tools without requiring extensive adaptation. Medina et al. [37] have pointed that the use of the same open data standards by public administrations can lead to greater interoperability. The Interoperability dimension helps to grade the extent to which data can be combined with other data and work together. The metrics depicted in Table 3.7 will be investigated. the metric Dataset is available in Open Format posses the highest score (from 0 to 70 out of possible points 100) due to Open format permits alter the data in any manner consider appropriate, without the necessity of possessing any proprietary software.

Table 3.7. Proposed metrics for Interoperability assessment

Metric	Description	Degree of Automation	Score (To-Be)
Dataset is available in Open Format	Dataset should be available in widely used and accessible formats (e.g., CSV, JSON, XML).	Manual	70
Number of formats available	Dataset available in multiple formats add flexibility to the Interoperability	Manual	10
Interoperability Maturity	The data should be accompanied by metadata, ensuring clear definitions and meanings of data elements	Manual	10
Licensing	The dataset should be accompanied by an open license	Manual	10

3.2.2.1. Metric : The dataset is available in Open Format

This measure the degree to which the dataset is provided in an Open Format. According to OpenDefinition [51], publicly available datasets should be provided in an open format. Open Formats are ones that can be processed with at least one open-source tool and whose use is unrestricted by either money or technology, in opposite, a proprietary format is usually controlled by a company. XML/JSON file types received the highest score because XML documents facilitate hierarchical structuring in a manner analogous to JSON, and in addition, the data format can frequently be effortlessly transformed into JSON through a variety of online tools [31]. The score of this metric is given according to the Table 3.8.

Table 3.8. File Format Score

Score	Format	Comment
0	PDF/TXT	Non-propietary format
15	XLS/XLSX/ZIP	Propietary format
35	CSV	Non-propietary format
70	XML/JSON	Non-propietary format

3.2.2.2. Metric : Number of formats available

The availability of the dataset in multiple formats enhances accessibility. The dataset should be provided in various formats to cater to different needs and tools. The scoring for this metric is depicted in Table 3.9.

Table 3.9. Number of formats published

Score	Formats Published	Comment
5	Only 1 format	The most common tabular formats is CSV
10	Multiple formats	Maximise reuse of the data

3.2.2.3. Metric : Interoperability Maturity

Dierickx [17] points out that the primary purpose of metadata is to document the characteristics of a dataset, thereby facilitating the subsequent re-utilization of the data. The explanation of data structures can improve interoperability due to the users's ability to interpret and reuse data correctly. The data should be accompanied by metadata,

ensuring clear definitions and meanings of data elements. Criteria and scoring for this metric is outlined in Table 3.10.

Table 3.10. Interoperability Maturity

Score	Criteria	Comment
0	Schema-less	Lack schema description
5	Schema	A schema does not explain the semantics of data; instead, it explains the syntax and structure.
10	Schema&Semantic Description	An explanation of each property/field of a data structure help users to interpreting and reusing data.

3.2.2.4. Metric : Licensing

The Licensing metric grade if the dataset is provided with an Open License.Rashid, Torchiano, et al. [56] state : *"publishing open data require the permission of the publisher, granted via an open licence"*. The dataset should be accompanied by an open license, permitting free use, modification, and distribution. According to Campos Zabala [8], when working with open data, it is critical to understand the license and legal implications associated with each dataset. Licenses specify the terms under which the data can be used, modified, and shared. Monino and Sedkaoui [40] propose that OD is distributed under an open license that ensures unrestricted access to it, allowing anyone to reuse it without encountering technical, legal, or financial constraints. Criteria and scoring for this metric is outlined in Table 3.11.

Table 3.11. Licensing score

Score	Criteria	Comment
0	No Open License attached to the dataset	The lack of a open license does not allow the reuse of the data
10	The dataset is provided with an open license	Open license permits improve reusability

3.2.3. Completeness dimension

The data is considered complete when it contains all necessary items for representing the entity. With very few exceptions (e.g. time-series data), the dataset should encompass all relevant facets of the topic it depicts. According to Mahanti [35], Completeness is the measure of whether data are present or absent. Sebastian-Coleman [62] states regarding Completeness of the dataset, the dataset should contain all of the records required by a data consumer, for a record to be complete, all mandatory fields must be populated, if a field is mandatory, then it should be fully populated (it should not contain any NULL). If a field is optional, then it should be populated or not based on the defined business rules.

The data is considered complete when it contains all necessary items for representing the entity. Completeness refers to the inclusion of all necessary information to adequately describe a concept at a satisfactory level of detail.

According to ISO [23], Completeness is the extent to which, under a particular usage context, subject data related to an entity includes values for all expected attributes

and related entity instances. Missing values compromise the level of completeness. In order to evaluate the Completeness dimension, the metric "Data is not null/missing in the dataset" is proposed. The metric illustrated in Table 3.12 will be calculated.

Table 3.12. Proposed metric for Completeness assessment

Metric	Description	Degree of Automation	Score (To-Be)
Data is not null/missing in the dataset	Define the quality criteria/fields to assessment	Automatic	25

3.2.3.1. Metric : Data is not null/missing in the dataset

The dataset should have minimal to no null values, ensuring the data is adequate to represent the entity. Open data sources might contain missing or incomplete data, potentially harming the efficiency of AI models. The absence of complete data could result in biased models or overfitting [8]. This metric is calculated in an automated way using Great Expectations (Table 3.13) evaluating each column selected by the data consumer (quality expectation) against the dataset evaluated.

Table 3.13. Mapping of Completeness dimension and Great Expectations

Score	Dimension	Metric	Great Expectation
25	Completeness	Data is not null/missing in the dataset	expect_column_values_to_not_be_null

The lines below Code Listing:3.1 show an extract of Python code developed of how expectation ("expect_column_values_to_not_be_null") will be applied to "Title" and "Authors" columns of the dataset .

Code Listing 3.1. Completeness using Great Expectations

```
# List of expectations for each column to be assess
# Expectation for columns "Title", "Author"
# mostly=1 parameter indicates values are never null, mostly=0.95 this
                                    would assert columns are null no more
                                     than 5%, mostly=0.97 this would
                                    assert columns are null no more than
                                    3%
#
columns_dataset = ["Title", "Authors"]
for column in columns_dataset:
    try:
        result = validator.expect_column_values_to_not_be_null(column=column
                                                , mostly=0.97, meta={"
                                                Dimension": "COMPLETENESS", "
                                                Metric": "Non-null/UnKnow data
                                                "})
```

```
        assert result.success , f"Column {column} have null values. Percent
                                                of Null Records {round(result.
                                                result['unexpected_percent'],4
                                                )}% ({result.result['
                                                unexpected_count']} of {result
                                                .result['element_count']}
                                                records have null value)"
        print(f"Column {column} do not have null values, unexpected_percent
                                                of records is : {result.result
                                                ['unexpected_percent']}  ")
    except AssertionError as e:
        print(e)
# Now we create a checkpoint to store all the data related to expectation
                                        validation
checkpoint_pre_fixes = gx.checkpoint.SimpleCheckpoint(name="
                                        checkpoint_bibliografia_pre_fixes",
                                        data_context=context,validations=[{"
                                        batch_request":
                                        bibliografia_batch_request,"
                                        expectation_suite_name":
                                        expectation_suite_name,},
    ],)
# Now we save our expectations
validator.save_expectation_suite(discard_failed_expectations=False)
# Now we run the checkpoint to validate the expectations against the data
                                        we extracted from the batch_request
checkpoint_result = checkpoint_pre_fixes.run()
#Generate a report using the DataDocs feature from Great Expectations
context.build_data_docs()
{'local_site': 'file://C:\\Users\\Abelardo\\AppData\\Local\\Temp\\
                                        tmp8gcagny8\\index.html'}
```

The results of the assessment are saved in a HTML report as show in the last line of the code above and presented the "Success Percent". In order to compute the points scored of the Completeness dimension, apply the "Success Percent" value to the Completeness dimension points (Table 3.12), e.g. for a 100% Success the dimension is scored with the maximum score (25), for other values of Success apply proportionally.

3.2.4. Uniqueness dimension

According to Zhang et al. [75], the concept of Uniqueness dictates that the data must possess an unique identification. For a dataset created from a relational database, the guarantee of uniqueness is achieved through the enforcement of primary key constraints. The examination of uniqueness concerns entails comparing the total number of data entries with the number of distinct values associated with a primary key attribute (whether existing or inferred). In datasets without ID or Primary Key,

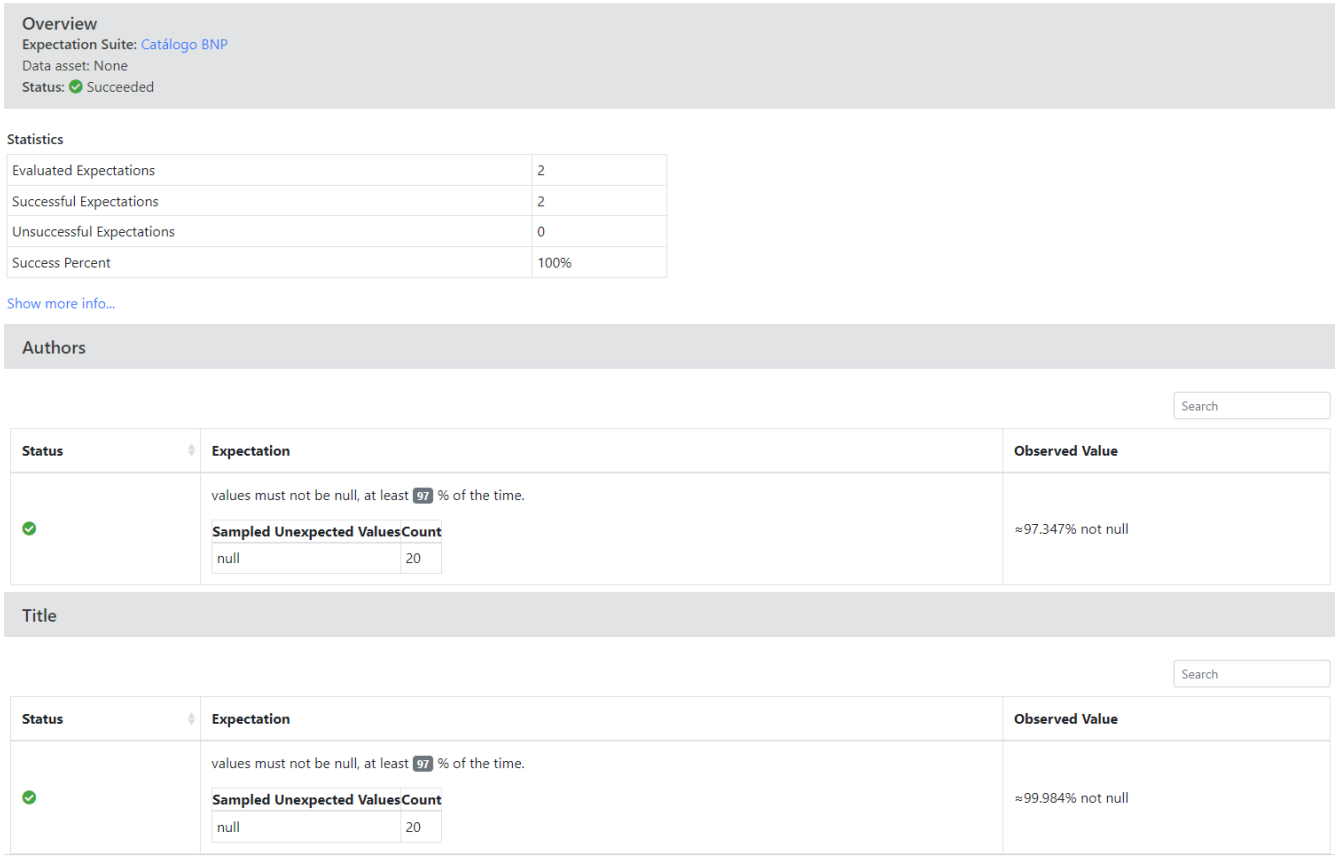


Figure 3.3. Completeness Assessment metric score using Great Expectations

the measure of uniqueness is determined by compound columns/fields assessment in order to identify repeated records. The proposed metric for Uniqueness assessment is outlined in Table 3.14.

Table 3.14. Proposed metric for Uniqueness assessment

Metric	Description	Degree of Automation	Points
Non-duplicates records	Identify key field or compound columns to be unique in order to assessment	Automatic	15

3.2.4.1. Metric : Non-duplicates records

Duplicated data occur when a dataset contains multiple copies of the same data or data records. Duplicate data is of no additional value. Instead, it lowers the quality of the data as it might cause errors during further processing [18]. This measure is calculated in an automated way using Great Expectations (Table 3.15) and computed based on a column chosen as primary ID (quality expectation) using the "expect_column_values_to_be_unique" expectation. In datasets provided without a key field or ID, the framework proposed assess uniqueness using compound columns using "expect_compound_columns_to_be_unique" expectation.

The lines below Code Listing :3.2 show an extract of Python code developed of how

Table 3.15. Mapping of Uniqueness dimension and Great Expectations

Score	Dimension	Metric	Great Expectations
15	Uniqueness	Non-duplicates records	expect_column_values_to_be_unique
			expect_compound_columns_to_be_unique

expectation ("expect_column_values_to_be_unique") will be applied to "BNP record ID" key column of the dataset.

Code Listing 3.2. Uniqueness using Great Expectations

```
.......
#
# Expectations for column "BNP record ID" should have no duplicate values
#
validator.expect_column_values_to_be_unique(column='BNP record ID', meta={"
                                        Dimension": "UNIQUENESS", "Metric": "
                                        Non-duplicate data"})
# Now we create a checkpoint to store all the data related to expectation
                                        validation
checkpoint_pre_fixes = gx.checkpoint.SimpleCheckpoint(name="
                                        checkpoint_bibliografia_pre_fixes",
                                        data_context=context,validations=[{"
                                        batch_request":
                                        bibliografia_batch_request,"
                                        expectation_suite_name":
                                        expectation_suite_name,},
    ],)
# Now we save our expectations
validator.save_expectation_suite(discard_failed_expectations=False)
# Now we run the checkpoint to validate the expectations against the data
                                        we extracted from the batch_request
checkpoint_result = checkpoint_pre_fixes.run()
#Generate a report using the DataDocs feature from Great Expectations
context.build_data_docs()
{'local_site': 'file://C:\\Users\\Abelardo\\AppData\\Local\\Temp\\
                                        tmpzco6psdf\\index.html'}
```

The results of the assessment are saved in a HTML report as show in the last line of the code and presented the "Success Percent" (Figure 3.4) . In order to compute the points scored of the Uniqueness dimension, apply the "Success Percent" value to the Uniqueness dimension points (Table 3.15), e.g. for a 100% Success the dimension is scored with the maximum score (15), for other values of Success apply proportionally.

In dataset without ID or primary key, the framework proposed use compound columns selected by the user in order to compute Uniqueness using the expectation "expect_compound_columns_to_be_unique" as show in Code Listing:3.3.

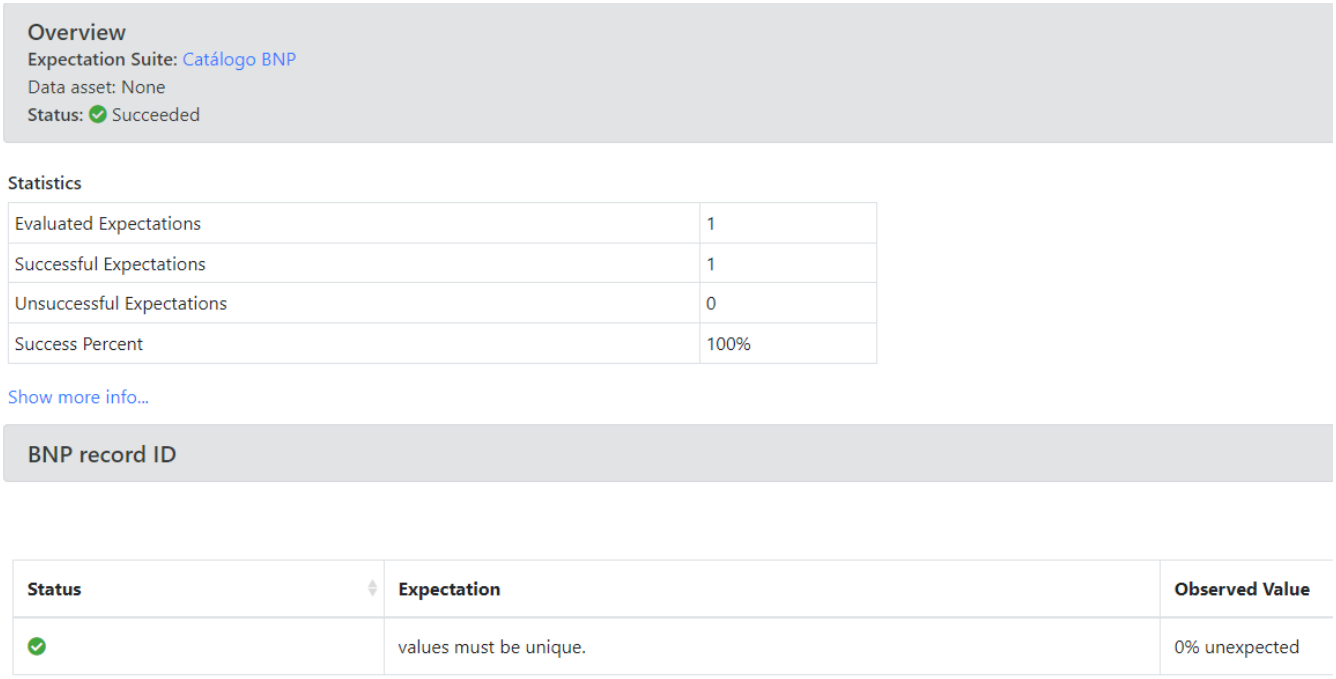


Figure 3.4. Uniqueness Assessment metric score using Great Expectations

Code Listing 3.3. Uniqueness using Great Expectations with multiples columns

```
# Expectations for column "Periodo","Regiao","Area CSP" must possess
                                    uniqueness collectively
validator.expect_compound_columns_to_be_unique(column_list = ["Periodo", "
                                    Regiao","Area CSP"], meta={"Dimension
                                    ": "UNIQUENESS", "Metric": "Non-
                                    duplicate data"})
```

An example of result of this assessment in depicted in the Figure 3.5.

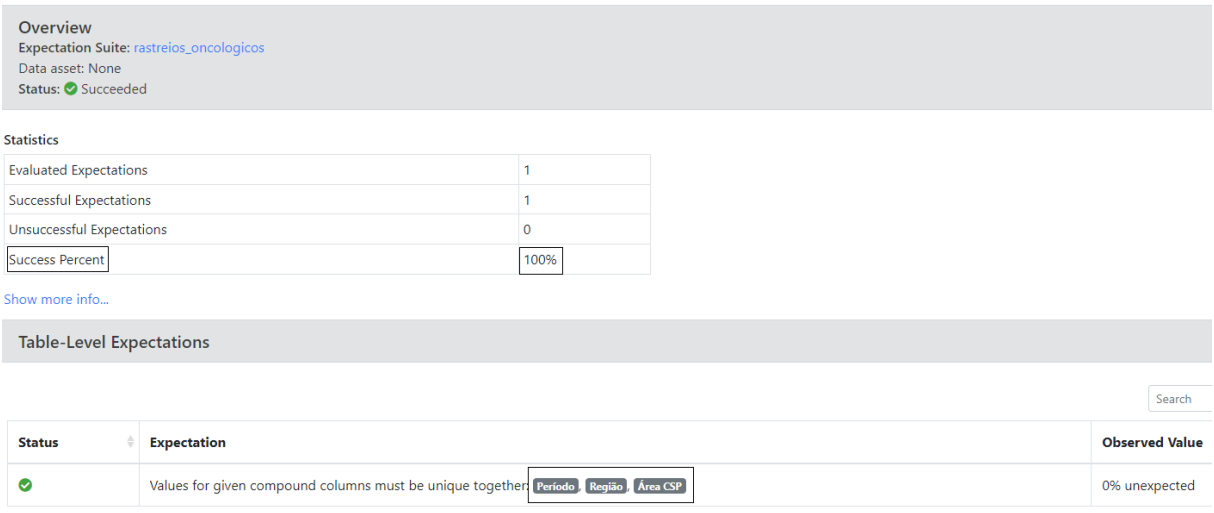


Figure 3.5. Uniqueness Assessment using Great Expectations - Compound Columns

3.3. Tools

The experiment environment was deployed as an application based on Python programming, Pandas and Great Expectations libraries. The provided Python code detects possible problems with data quality for the Completeness and Uniqueness dimensions.

Python[1], is a programming language extensively utilized in web applications, software development, data science, and machine learning (ML). Python software is downloadable for free, seamlessly integrates with diverse systems, and accelerates the pace of development.

Pandas[2], is a specialized Python library for managing, analyzing, and processing data. It relies on the data structures of the NumPy library, making it a prerequisite when installing Pandas. Within this library, three structures are at your disposal: Series, DataFrame, and Panel.

Great Expectations, is an open-source tool built in Python. It has several major features including data validation, profiling, and documenting data projects.

How it will works:

(1) Preview the dataset (if the open portal enables the functionality);
(2) Investigate about domain of the dataset in order to identify data quality requirements;
(3) Data Exploration using Pandas;
(4) Set up quality conditions and check using Great Expectations based on previous step (domain of dataset). The quality conditions must be based on / related with quality dimensions;
(5) Calculate the quality score.

An schematic technical diagram of the framework is depicted in the Figure 3.6.

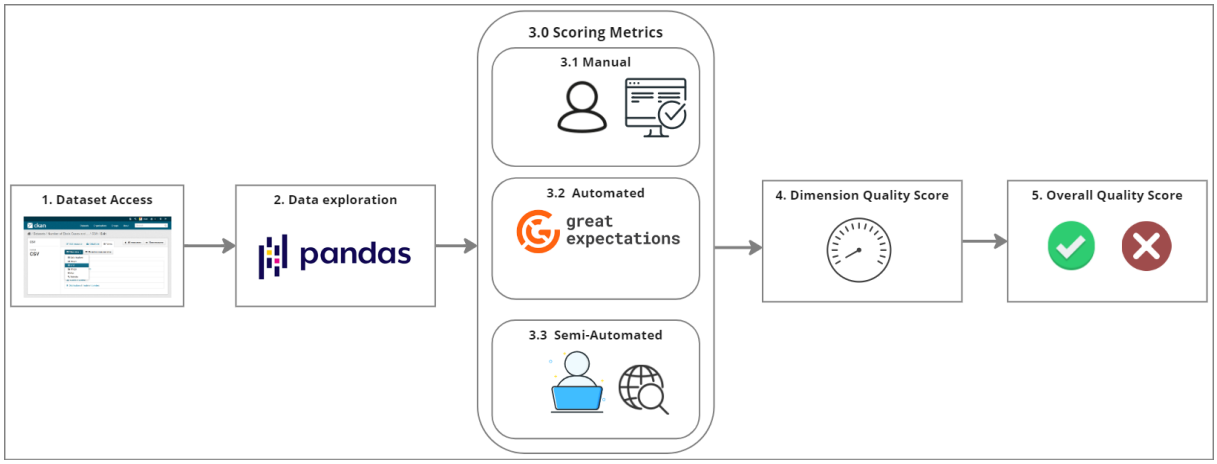


Figure 3.6. Technical diagram of the proposed framework
Source : By the Author

[1]https://www.python.org/
[2]https://pandas.pydata.org/

CHAPTER 4

Demonstration: Open Data Quality Assessment

This section illustrates the process of open data quality evaluation through the framework presented in the Chapter 3. According to the European Union et al. [19], Open Data that is provided by governmental entities for the purpose of reuse should be officially released and easily reachable by those who wish to reuse it. Regarding this matter, Open Data Portals are online platforms established by government agencies to act as repositories that facilitate the exploration of public data assets.

Metrics for Completeness and Accuracy were computed automatically using Python and Great Expectations applying expectation(s) defined in Table 3.13 for Completeness and Table 3.15 for Uniqueness. It was necessary to manually or semi-automated way calculate Accessibility and Interoperability and associated metrics as defined in Figure 3.1. This assessment focuses in the content of the open datasets in terms of the framework proposed.

According to Mateus [36], there are a total of 12 distinct open data portals in Portugal. The Framework proposed was validated through eight datasets downloaded from different portuguese Open Data Portals as outlined in Table 4.1.

Table 4.1. Portuguese Open Datasets

Portal Name	Dataset Name	Description	Source/Data Producer	Records
dados.gov.pt	Condicionamentos de Trânsito	Active and planned traffic restrictions in the city of Lisbon. Provides information on the location, reason, type of restriction and impact on traffic	Câmara Municipal de Lisboa	3,779
	Estabelecimentos de Alojamento Local	Local lodging establishments in Portugal	Turismo de Portugal I.P.	114,022
	Justiça no mapa	Georeferenced geographical location of facilities linked to the Ministry of Justice in Portugal	Direção-Geral da Política de Justiça	2,211
	Estrutura de Missão PRR - Contratualização	PRR contractualisation data	Agência para o Desenvolvimento e Coesão, IP	190
transparencia.sns.gov.pt	Rastreios Oncológicos	Monitoring the cancer screening programme in primary health care.	Central Administration of the Health System, I.P.	6,131
	Saúde Oral	Activity of the National Oral Health Promotion Programme through the "Cheque Dentista" instrument.	SISO - Oral Health Information System	1,372
opendata.bnportugal.gov.pt	Bibliografia Nacional Portuguesa	Bibliographic records of monographs and serial publications published in Portugal since 2002	Biblioteca Nacional de Portugal - BNP	294,020
data.cascais.pt	Edifícios de Habitação Municipal	Register of municipal housing buildings intended for social housing and their characterisation	Câmara Municipal de Cascais	590
	Rastreios Oncológicos	Monitoring the cancer screening programme in primary health care	SPMS– Serviços Partilhados do Ministério da Saúde	108

The datasets are from different data producers, which permits the examination of discrepancies in data quality between data sourced from governmental entities and those from municipal sources. According to Open Data Maturity Report 2023[1], Portugal's open data quality score is 89%, therefore, it is expected that selected datasets provide high quality data.

[1]https://data.europa.eu/en/publications/open-data-maturity/2023

4.1. Dataset - Condicionamentos de Trânsito Lisboa

The Lisbon Municipality provides a dataset for active and planned traffic restrictions. It has 3,961 records and nine columns as depicted in Table 4.2.

Table 4.2. Condicionamentos de Trânsito Lisboa - Dataset

Order	Field Name	Data Type	Comment
1	creation_date	DateTime	Date registered
2	entity_id	String	Conditioning ID
3	position	String	Coordinates (lat/long)
4	impacto	String	Type of impact on traffic
5	morada	String	Conditioning address
6	motivo	String	Reason for conditioning
7	periodos_condicionamentos	String	Start date and End date of conditioning
8	restricao_circulacao	String	Type of traffic restriction
9	periodos_condicionamentos['date_max']	String	N/A

Source: https://dados.gov.pt/pt/datasets/condicionamentos-de-transito/

Dimension : Accessibility

- **Accessible for download** :
  This metric refers to the fact that the dataset can be accessed by a human being or a computer without encountering any access restrictions. Download dataset from the open data portal was downloaded manually in a single file. There is no option for consuming via API. The score achieved was 40 points out of 60 possible points;
- **Registration is necessary** :
  For data consumers, no previous registration is needed in the Open Data Portal, this allow easy consumption and further processing of the dataset. A dataset was obtained directly by clicking on the 'download' icon. The score achieved was 15 points (maximum score);
- **Download URL is accessible** :
  To evaluate this metric, we used the HTTP status check tool available at https://httpstatus.io/, which takes the URL as a parameter. This tool can be used to manually check whether the connection downloading your data is accessible or not. The URL for downloading the dataset was working and accessible, its up-time is 100%. The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available in "Comma-Separated Value" (.CSV). The data is accessible in a structured and non-proprietary format. According to the framework, datasets possessing a tabular format, such as CSV files, are awarded 35 points out of 70 possible points;

- **Number of formats available** :
  Availability of the dataset in multiple formats enhances interoperability. The Lisbon Municipality have published the dataset in one file format "Comma-Separated Value" (CSV). The score achieved was 5 points out of 10 possible points;
- **Interoperability Maturity** :
  This metric enables us to evaluate the presence of an explanation for each property or field within a data structure, assisting users in interpreting and reusing the data. It was identified only data about the schema in the source/data provider. The score achieved was 5 points out of 10 possible points;
- **Licensing** :
  In terms Licensing, the dataset is provided with an Open License (Creative Commons Attribution 4.0 - CC BY 4.0). The score achieved was 10 points (maximum score);

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  To calculate the respective metrics, we have developed a list of all the fields in the dataset within the Python code, and for each of them assessed their quality using the Great Expectations library. The score obtained was 25 points (maximum score).

Dimension : Uniqueness

- **Non-duplicates records** :
  We detected a high numbers of duplicated records. For the dataset being analysed, we used the 'entity_id' field to identify each record unequivocally. The score obtained was 0.45 points out of 15 possible points. In order to confirm this low value, a double check was performed comparing all columns of the dataset obtaining the same value.

The result of the assessment reveals a quality score of 65.50 (out of 100), which is less than the Portugal Quality Score[2] (**89**%). Some quality problems were detected impacting in the final quality score of the dataset as outlined in the Table 4.3.

The dimension with low quality score (0.45 out of 15) is Uniqueness due to duplicate rows published in the dataset. When dealing with uniqueness, the identifier attribute ("entity_id") for the assessed dataset help us to cope with the possibility of duplicate rows. Additionally, the dataset was assessed using all columns obtaining the same result

Other quality dimensions as Accessibility (30 out of 30), Interoperability (16.5 out of 30) and Completeness (25 out of 25) present scores expected.

[2]https://data.europa.eu/en/publications/open-data-maturity/2023

Table 4.3. Condicionamentos de Trânsito Lisboa - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**40**	**12**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**35**	**10.5**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**5**	**1.5**
		Licensing	10	**10**	**3**
Completeness	25	Data is not null/missing in the dataset	100	**100**	**25**
Uniqueness	15	Non-duplicates records	100	**0.03**	**0.0045**
				Quality Score	**65.50**

Source: By the author

4.2. Dataset - Estabelecimentos de Alojamento Local

The Turismo de Portugal I.P. provides a register of Local Accommodation Establishments, the dataset has 50 fields and 114,022 depicted in Table 4.4.

Dimension : Accessibility

- **Accessible for download** :
  Data is accessible for the user. The dataset was downloaded manually, which took a long time because there is only one file with aggregated data (the number or records is about 114,00 records) and the data is generated in real-time, therefore 30 points out of 60 possible points were assigned;
- **Registration is necessary** :
  No prior registration on the Open Data Portal was required to download the dataset, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  No errors were returned, data is accessible for the user, but due to the size of the file, the message returned is : "the request has been accepted for processing, but the processing has not been finished yet". The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available different open formats. Regarding accessible in a structured and non-proprietary format, dataset is available in "Comma-Separated Value" (.CSV). Metric was awarded with 35 points out of 70 possible points;
- **Number of formats available** :
  Availability of the dataset in multiple formats enhances interoperability. The open portal have published the dataset in four file formats (CSV, GeoJSON, KML and Shapefile). According to the framework, JSON/XML file types received the highest score. The score achieved was 10 points (maximum score);
- **Interoperability Maturity** :
  This measure evaluates the availability of explanations for each property or

Table 4.4. Estabelecimentos de Alojamento Local - Dataset

Order	Field Name	Data Type	Comment
1	X	DateTime	N/A No Info
2	Y	String	N/A No Info
3	ObjectID	Number	Autonumber ID
4	DataRegisto	Date	Date registered
5	DataAberturaPublico	Date	Public opening date
6	Denominacao	String	Designation/Name
7	Modalidade	String	Modality/Type
8	NrUtentes	String	Nº guests
9	TitularExploracao	String	Owner's name
10	Endereco	String	Address
11	CodigoPostal	String	ZIP Code
12	LOCALIDADE	String	Postcode Town
13	ObservacoesTdP	String	Observations
14	LatLong	String	Latitude, Longitude
15	FiabilidadeGeo	String	Reliability of georeferencing
16	Freguesia	String	Neighbourhood
17	Concelho	String	County
18	Distrito	String	District
19	NUTSIII	String	NUTS III
20	NUTSII	String	NUTSII
21	NUTSIICCDR	String	NUTSII CCDR
22	ERT	String	Regional Tourism Authority (ERT)
23	NUTSI	String	NUTS I (Continent)
24	RNAP	String	National Network of Protected Areas
25	RedeNatura2000Global	String	Natura Network 2000 (Global)
26	RedeNatura2000Sitios	String	Natura Network 2000 - Sites
27	RedeNatura2000ZPE	String	Natura Network 2000 - ZPE
28	FaixaCosteira	String	Coastline (5km)
29	ZonaCosteira	String	Coastal Zone (2km)
30	ZonaTerrestreProtecao	String	Protected Land Zone (500m)
31	LimitePROT	String	PROT (Regional Land Management Plans)
32	LimitePOC	String	Coastline programmes (not exhaustive)
33	LimiteAlbufeiras	String	Reservoir programmes (not exhaustive)
34	UNESCO	String	UNESCO World Heritage Sites
35	Editor	String	Editor
36	DataEdicao	Date	Issue Date
37	Email	String	Email
38	Telefone	String	Telefone
39	Telemovel	String	Telefone
40	Geoparques	String	Geoparks
41	Biosfera	String	Biosphere Reserve
42	ConcessaoTuristica	String	Tourist Concession
43	NrRNAL	Integer	NrRNAL
44	PNCT	String	National Program for Territorial Cohesion
45	concelhoRNAL	String	County RNAL
46	EstacaoNautica	String	Nautical Station
47	RedeAldeias	String	Village Network
48	FreguesiasCosteiras	String	Coastal villages
49	NUTSIIICosteiras	String	NUTS III Coastline (EUROSTAT)
50	SeloCleanSafe	String	Clean & Safe label

Source: https://dados.gov.pt/pt/datasets/estabelecimentos-de-alojamento-local-1/

field in a data structure, helping users analyze and reuse data. The dataset only contain basic schema-related information in the source provider instead of the open data portal. The score achieved was 5 points out of 10 possible points;

- **Licensing** :
  License that has not been specified, this will limit free use, modification; and distribution. It is essential to comply with these licenses to avoid legal issues

[8]. It is important because if the dataset do not explicitly have an open license, it is not considered OD. The score achieved was 0 points out of 10 possible points;

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  In order to calculate the respective metrics, a list of all the fields in the dataset was defined (excluding those that are not in the metadata or are optional) based on the minimal meta-data documentation available on the source provider. For each of them (47), their quality was assessed, 11 were unsuccessful. The final result for the metric was 19 points out of 25 possible points.

Dimension : Uniqueness

- **Non-duplicates records** :
  For the dataset being analysed, we used the 'OBJECTID' as a primary key. Due to the dataset store geographical data, looks like this key is an artificial, self-increasing key. The score obtained was 15 points (maximum score).

The result of the assessment show that the quality score (**70.02**) is less than the Portugal Quality Score[3] (**89**%). Some quality problems were detected impacting in the final quality score of the dataset as outlined in the Table 4.7.

The dimension with a lower quality score (15 out of 30) is Interoperability due to limited structured open data formats and dataset without open license.

Quality dimensions as Accessibility (21 out of 30), Uniqueness (15 out of 15) and Completeness (19 out of 25) present high quality.

Table 4.5. Estabelecimentos de Alojamento Local - Quality Score

Dimension	Points	Metric	Metric Score (To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**30**	**9**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**35**	**10.5**
		Number of formats available	10	**10**	**3**
		Interoperability Maturity	10	**5**	**1.5**
		Licensing	10	**0**	**0**
Completeness	25	Data is not null/missing in the dataset	100	**76.09**	**19.02**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**70.02**

Source: By the author

4.3. Dataset - Justiça no mapa

The Direção-Geral da Política de Justiça (institution of the ministry of justice) provides a dataset Georeferenced geographical location of equipment linked to the Portuguese Ministry of Justice. It has 2,211 records and 19 columns as depicted in Table 4.8.

Dimension : Accessibility

[3]https://data.europa.eu/en/publications/open-data-maturity/2023

Table 4.6. Justiça no mapa - Dataset

Order	Field Name	Data Type	Comment
1	Nome	String	N/A
2	Tipo	Number	1 - tribunais, 2 - registos e notariado, 3 - centros de arbitragem, 4 - estabelecimentos prisionais, 5 - instituto nacional de medicina legal, 7 - julgados de paz.
3	Ordenacao	Number	N/A
4	Morada	String	N/A
5	CodigoPostal	String	N/A
6	Telefone	String	N/A
7	FAX	String	N/A
8	Email	String	N/A
9	Posicao/Lat	Number	N/A
10	Posicao/Lng	Number	N/A
11	DadosAtividade/0/Texto	String	N/A
12	DadosAtividade/0/Valor	Number	N/A
13	DadosAtividade/1/Texto	String	N/A
14	DadosAtividade/1/Valor	Number	N/A
15	DadosAtividade/2/Texto	String	N/A
16	DadosAtividade/2/Valor	Number	N/A
17	DadosAtividade/3/Texto	String	N/A
18	DadosAtividade/3/Valor	Number	N/A
19	DadosAtividade/0	String	N/A

Source: https://dados.gov.pt/pt/datasets/justica-no-mapa/

- **Accessible for download** :
  Data is accessible for the user. The dataset was downloaded manually and is not accessible via API. The score of was 40 points out of 60 possible points.;
- **Registration is necessary** :
  No prior registration on the Open Data Portal was required to download the dataset, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  No errors were returned, the status code was '200 OK' . The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available in an open format. Regarding accessible in a structured and non-proprietary format, dataset is available in "JavaScript Object Notation" (.JSON). Metric was awarded 70 points (maximum score);
- **Number of formats available** :
  Availability of the dataset in multiple formats enhances interoperability. The open portal have published the dataset in one file format (JSON). The score achieved was 5 points out of 10 possible points;
- **Interoperability Maturity** :
  This measure evaluates the availability of explanations for each property or

field in a data structure, helping users analyze and reuse data. The dataset has a basic description enumerating the name of the fields but not metadata is included. The score achieved was 5 points out of 10 possible points;

- **Licensing** :
  The dataset is provided with an Open License (Creative Commons Attribution 4.0 - CC BY 4.0). The score achieved was 10 points (maximum score);

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  In order to calculate the respective metrics, the dataset's fields Nome, Tipo, Ordenacao, Morada, CodigoPostal, Telefone, Email, Posicao/Lat, Posicao/Lng were selected (excluding those that are optional or identified to be null during the exploration initial of dataset and domain). For each of them (9), their quality was assessed, 3 were unsuccessful. The final result for the metric was 16.5 points out of 25 possible points.

Dimension : Uniqueness

- **Non-duplicates records** :
  Due to the dataset store geographical data presented in a structured format, there is not a key defined by the data producers during publish data. The framework proposed permits assess Uniqueness using the expectation "expect_compound_columns_to_be_unique" (checks that every combination of values in the specified columns is unique across all rows in the dataset) proposed in Table 3.15. This expectation was applied to the columns 'Nome', 'Tipo', 'Ordenacao' and 'Morada'. The score obtained was 15 points (maximum score).

The result of the assessment reveals a quality score of 82.50 (out of 100), which is less than the Portugal Quality Score[4] (**89**%). Some quality problems were detected impacting in the final quality score of the dataset as outlined in the Table 4.7.

Table 4.7. Justiça no mapa - Quality Score

Dimension	Points	Metric	Metric Score (To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**40**	**12**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**70**	**21**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**0**	**0**
		Licensing	10	**10**	**3**
Completeness	25	Data is not null/missing in the dataset	100	**66**	**16.50**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**82.50**

Source: By the author

[4]https://data.europa.eu/en/publications/open-data-maturity/2023

4.4. Dataset - PRR - Contratualização

The Portugal government provides a dataset for PRR[5] contractualisation data. It has 190 records and 13 columns as depicted in Table 4.8.

Dimension : Accessibility

Table 4.8. PRR - Contratualização - Dataset

Order	Field Name	Data Type	Comment
1	Data_date	Date	N/A
2	Código do Contrato	String	N/A
3	Designação do Contrato	String	N/A
4	Data da Assinatura do Contrato	Date	N/A
5	Montante Contratualizado	Number	N/A
6	Montante subvenção	Number	N/A
7	Montante empréstimos	Number	N/A
8	Código do investimento	String	N/A
9	Código da Entidade	String	N/A
10	Designação da Entidade	String	N/A
11	Estado Contratualização	String	N/A
12	Valor Transferido	Number	N/A
13	Valor Pago	Number	N/A

Source: https://dados.gov.pt/pt/datasets/dataset-estrutura-de-missao-prr-contratualizacao/

- **Accessible for download** :
  The dataset was manually download quickly due to does not come from an external portal. API for this dataset is not available for data consuming. The score of 40 points out of 60 points was assigned;
- **Registration is necessary** :
  Access to the dataset was not limited due to a prior registration in the Open Data Portal to download the dataset, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  Using https://httpstatus.io/ tool, we validate accessibility to the dataset using URLs (2) available in the open data portal. The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is not available open format. Regarding accessible in a structured and non-proprietary format, dataset is only available in "Excel" (.XLSX), due to is a machine-readable, metric was awarded 15 points out of 70 possible points;
- **Number of formats available** :
  Availability of the dataset in multiple formats enhances interoperability. The

[5]https://recuperarportugal.gov.pt/

open portal have published the dataset only in EXCEL (.XLSX), this limit interoperability and minimize the reuse of dataset. The score achieved was 5 points out of 10 possible points;

- **Interoperability Maturity** :
  This measure evaluates the availability of explanations for each property or field in a data structure, helping users analyze and reuse data. The dataset is schema-less, there is not any kind of information about the fields or related metadata. The score achieved was 0 points out of 10 possible points;
- **Licensing** :
  The dataset is provided with an Open License (Creative Commons Attribution 4.0 - CC BY 4.0). The score achieved was 10 points (maximum score);

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  Regarding this metric, we assess all fields of the dataset (13) with success. The final result for the metric was 25 points (maximum score).

Dimension : Uniqueness

- **Non-duplicates records** :
  The dataset have a key field "Código do Contrato", this field was used for assessment obtaining as score of 15 points (maximum score).

The result of the assessment reveals a quality score of 73 (out of 100), which is less than the Portugal Quality Score[6] (**89**%). Some quality problems were detected impacting in the final quality score of the dataset as outlined in the Table 4.9.

Table 4.9. PRR - Contratualização - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**40**	**12**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**15**	**4.5**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**0**	**0**
		Licensing	10	**10**	**3**
Completeness	25	Data is not null/missing in the dataset	100	**100**	**25**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**73.00**

Source: By the author

4.5. Dataset - Rastreios Oncológicos (SNS)

The National Health System (SNS) publish a dataset with data about the Monitoring the cancer screening programme in primary health care. The dataset has 6,131 records and 10 columns as depicted in Table 4.12.

[6]https://data.europa.eu/en/publications/open-data-maturity/2023

Table 4.10. Rastreios Oncológicos (SNS) - Dataset

Order	Field Name	DataType	Comment
1	Período	Date	N/A
2	Região	String	N/A
3	Área CSP	String	N/A
4	Localização Geográfica	String	Type=geo_point_2d
5	Mulheres com Registo de Mamografia nos Últimos Dois Anos	Number	N/A
6	Mulheres [50; 70[A, c/ mamogr. (2 anos)	Number	N/A
7	Mulheres com Colpocitologia Atualizada	Number	N/A
8	Mulheres [25; 60[A, c/ colpoc. Atualizada	Number	N/A
9	Utentes Inscritos com Rastreio do Cancro do Colon e Reto Efetuado	Number	N/A
10	Utentes [50; 75[A, c/ rastreio cancro CR	Number	N/A

Source: https://transparencia.sns.gov.pt/explore/dataset/rastreios-oncologicos/

Dimension : Accessibility

- **Accessible for download** :
  The dataset was manually download without any restriction. Additionally , the dataset can be consumed via an API that allows to search and download records using various parameters. The score of 60 points was assigned (maximum score);
- **Registration is necessary** :
  Access to the dataset does not require prior registration in the Open Data Portal to download the dataset or consuming via API, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  The accessibility to the dataset using URL provided was validated with success. The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available in two open format (CSV,JSON). The metric was awarded 70 points (maximum score);
- **Number of formats available** :
  The open portal have published the dataset three data format (CSV, JSON and EXCEL). Additionally the dataset is published in Geographic file formats (GeoJSON, Shapefile, KML and GPX), Data analysis file formats (Parquet). The score achieved was 10 points (maximum score);
- **Interoperability Maturity** :
  This measure evaluates the availability of explanations for each property or field in a data structure, helping users analyze and reuse data. The dataset has schema with basic information about the fields and metadata.The score achieved was 5 points out of 10 possible points;

- **Licensing** :
  Licence not specified. The score achieved was 0 point out of 10 possible points;

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  Regarding this metric, is expected the values must not be null due to the preview of a group of records of the dataset in the available in the open data portal. The framework assessed all fields of the dataset (10) with success. The final result for the metric was 25 points (maximum score).

Dimension : Uniqueness

- **Non-duplicates records** :
  The dataset does not have a key field. In order to assess Uniqueness dimension, the values corresponding to the specified compound columns ('Período', 'Região', 'Área CSP') must possess uniqueness collectively. No duplicate records were identified using the expectation "expect_compound_columns_to_be_unique", therefore, the result was 15 points (maximum score).

The result of the assessment reveals a quality score of 95.50 (out of 100). This score represent a dataset with high data quality when comparing to Portugal Quality Score[7] (**89**%). The detailed dimensions ans metrics score is outlined in Table 4.11.

Table 4.11. Rastreios Oncológicos (SNS) - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**60**	**18**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**70**	**21**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**10**	**3**
		Licensing	10	**0**	**0**
Completeness	25	Data is not null/missing in the dataset	100	**100**	**25**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**95.50**

Source: By the author

4.6. Dataset - Saúde Oral (SNS)

The National Health System (SNS) publish a dataset with data about the Cheques and referrals (SOCSP and HO) issued, used and treatments carried out. The dataset has 1,372 records and 8 columns as depicted in Table 4.12.

Dimension : Accessibility

[7] https://data.europa.eu/en/publications/open-data-maturity/2023

Table 4.12. Saúde Oral (SNS) - Dataset

Order	Field Name	DataType	Comment
1	Período	Date	N/A
2	Entidade	String	N/A
3	População-Alvo	String	N/A
4	Âmbito de Intervenção	String	N/A
5	Nº Cheques Emitidos	Number	N/A
6	Nº Cheques Utilizados	Number	N/A
7	Nº Tratamentos Efetuados	Number	N/A
8	ID	String	N/A

Source: https://transparencia.sns.gov.pt/explore/dataset/saude-oral

- **Accessible for download** :
  The dataset was manually download without any restriction. Additionally , the dataset can be consumed via an API that allows to search and download records using various parameters. The score of 60 points was assigned (maximum score);
- **Registration is necessary** :
  Access to the dataset does not require prior registration in the Open Data Portal to download the dataset or consuming via API, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  The accessibility to the dataset using URL provided was validated with success. The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available in two open format (CSV,JSON). The metric was awarded 70 points (maximum score);
- **Number of formats available** :
  The open portal have published the dataset three data format (CSV, JSON and EXCEL). The score achieved was 10 points (maximum score);
- **Interoperability Maturity** :
  The dataset has schema with basic information about the fields and meta-data.The score achieved was 5 points out of 10 possible points;
- **Licensing** :
  Licence not specified. The score achieved was 0 point out of 10 possible points;

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  Regarding this metric, is expected the values must not be null due to the preview of a group of records of the dataset in the available in the open data portal. The framework assessed all fields of the dataset (8) with success. The final result for the metric was 25 points (maximum score).

Dimension : Uniqueness

- **Non-duplicates records** :
  The dataset have a key field named "ID". No duplicate records were identified using the expectation "expect_column_values_to_be_unique", therefore, the result was 15 points (maximum score).

The result of the assessment reveals a quality score of 95.50 (out of 100). This score represent a dataset with high data quality when comparing to Portugal Quality Score[8] (**89**%). The detailed dimensions ans metrics score is outlined in Table 4.13.

Table 4.13. Saúde Oral (SNS) - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**60**	**18**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**70**	**21**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**10**	**3**
		Licensing	10	**0**	**0**
Completeness	25	Data is not null/missing in the dataset	100	**100**	**25**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**95.50**

Source: By the author

4.7. Dataset - Catálogo BNP - Portugal

The National Library of Portugal (BNP) provides a dataset with the entire bibliographic catalogue. The dataset has 1'349,062 records and 21 columns as depicted in Table 4.14.

Dimension : Accessibility

- **Accessible for download** :
  The dataset was manually download manually and is not available for consuming via an API. This dataset is published in one file due to contains master data of bibliographic records. The score achieved was 40 points out of 60 possible points;
- **Registration is necessary** :
  Access to the dataset does not require prior registration, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  The accessibility to the dataset through the provided URL was successfully verified. The score achieved was 25 points (maximum score);

Dimension : Interoperability

[8]https://data.europa.eu/en/publications/open-data-maturity/2023

Table 4.14. Catálogo BNP - Dataset

Order	Field Name	DataType	Comment
1	BNP record ID	Number	N/A
2	Material type	String	N/A
3	ISBN	String	N/A
4	Legal deposit number	String	N/A
5	Language of Text	String	N/A
6	Language of Original Work	String	N/A
7	Title	String	N/A
8	Subtitle	String	N/A
9	Original title	String	N/A
10	Edition	String	N/A
11	Place of publicattion	String	N/A
12	Name of Publisher	String	N/A
13	Date of Publication	String	N/A
14	Extent of Item	String	N/A
15	Dimensions	String	N/A
16	Series	String	N/A
17	Volume	String	N/A
18	Universal Decimal Classification	String	N/A
19	Authors	String	N/A
20	Image	String	N/A
21	Persistent URL	String	N/A

Source: https://opendata.bnportugal.gov.pt/docs/catalogo.csv.zip

- **Dataset is available in Open Format** :
  The dataset is available compressed in a ZIP file. The metric was awarded 15 points out of of 70 points;
- **Number of formats available** :
  The open portal have published the dataset in one data format (ZIP). The score achieved was 5 points out of 10 possible points;;
- **Interoperability Maturity** :
  The dataset has no any information about the fields and metadata. The score achieved was 0 points out of 10 possible points;
- **Licensing** :
  The dataset is provided under the license CC0 (Creative Commons CC0 1.0 Universal Public Domain Dedication),so 10 points were awarded (maximum score);

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  Regarding this metric, is expected the values must not be null due to the preview of a group of records of the dataset in the available in the open data portal. The framework assessed two principal fields ('Title', 'Authors') with partial success. The score achieved was 24 points out of 25 possible points;.

Dimension : Uniqueness

- **Non-duplicates records** :
  The dataset have a key field named "BNP record ID". No duplicate records were identified, the result was 15 points (maximum score).

The result of the assessment reveals a quality score of 72 (out of 100). This score represent a dataset with low data quality when comparing to Portugal Quality Score[9] (**89**%). The detailed dimensions ans metrics score is outlined in Table 4.15.

Table 4.15. Catálogo BNP - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**40**	**12**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**15**	**4.5**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**0**	**0**
		Licensing	10	**10**	**3**
Completeness	25	Data is not null/missing in the dataset	100	**98**	**24**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**72**

Source: By the author

4.8. Dataset - Edifícios de Habitação Municipal Cascais

The Cascais Municipality provides a dataset for Register of municipal housing buildings intended for social housing and their characterisation. It has 590 records and 11 columns as depicted in Table 4.16.

Table 4.16. Edifícios de Habitação Municipal Cascais - Dataset

Order	Field Name	DataType	Comment
1	Id	Date	N/A
2	Tipo de Utilização	String	N/A
3	Rua	String	N/A
4	Local	String	N/A
5	Freguesia	String	N/A
6	Número de polícia	String	N/A
7	Total de Pisos	Number	N/A
8	Número de Fogos da CMC	Number	N/A
9	Número de Fogos Total	Number	N/A
10	Área Implantação	String	N/A
11	Data de actualização	Date	N/A

Source: https://data.cascais.pt/geral/data-hub/dados-abertos

Dimension : Accessibility

- **Accessible for download** :
  The dataset was manually download manually and is not available for consuming via an API. The score achieved was 40 points out of 60 possible points;

[9] https://data.europa.eu/en/publications/open-data-maturity/2023

- **Registration is necessary** :
  Access to the dataset does not require prior registration, so 15 points were awarded (maximum score);
- **Download URL is accessible** :
  The accessibility to the dataset through the provided URL was successfully verified. The score achieved was 25 points (maximum score);

Dimension : Interoperability

- **Dataset is available in Open Format** :
  The dataset is available in CSV format. The score achieved was 35 points out of 70 possible points;
- **Number of formats available** :
  The dataset is available in only one structured tabular format (CSV). The dataset is published in additional no structured format (geojson, kmz, shp). The score achieved was 5 points out of 10 possible points;
- **Interoperability Maturity** :
  The dataset has no any information about the fields and metadata. The score achieved was 0 points out of 10 possible points;
- **Licensing** :
  The dataset is provided under the license Creative Commons - Domínio Público ,so 10 points were awarded (maximum score);

Dimension : Completeness

- **Data is not null/missing in the dataset** :
  The framework assessed one principal field ('Total de Pisos') with partial success. The score achieved was 8 points out of 25 possible points;

Dimension : Uniqueness

- **Non-duplicates records** :
  The dataset have a key field named "ID". No duplicate records were identified, the result was 15 points (maximum score).

The detailed dimensions ans metrics score is outlined in Table 4.17.

Table 4.17. Edifícios de Habitação Municipal Cascais - Quality Score

Dimension	Points	Metric	Metric Score(To-Be)	Metric Scored	Points Scored
Accessibility	30	Accessible for download	60	**40**	**12**
		Registration is necessary	15	**15**	**4.5**
		Download URL is accessible	25	**25**	**7.5**
Interoperability	30	The dataset is available in Open Format	70	**35**	**10.5**
		Number of formats available	10	**5**	**1.5**
		Interoperability Maturity	10	**0**	**0**
		Licensing	10	**10**	**3**
Completeness	25	Data is not null/missing in the dataset	100	**31**	**7.75**
Uniqueness	15	Non-duplicates records	100	**100**	**15**
				Quality Score	**61.75**

Source: By the author

CHAPTER 5

Results analysis

As previously said, the main goal of this dissertation is to suggest a framework for assessing the quality of open data. Regarding the quality dimensions and metrics used for assess open data (RQ1), different authors consider important to define some dimensions of quality. Based on the review of the literature, the dimensions that are predominantly cited by the authors are Completeness, Accuracy, Consistency, Understandability, Accessibility, Timelines and Uniqueness. Some authors define the same concept but with different name which generates ambiguity (e.g. Timeliness and Freshness). A dimension of data quality consists of one or more metrics or criteria. The most used metrics are related to licensing information , column completeness, file format, cell with correct values (according the domain) and duplicate rows.

There is no common framework used in order to assess open data. The most used framework for assess Open Data (RQ2) is 5-Start Open Data rating system proposed by Tim Berners-Lee. The model establishes five levels of quality for open data with a focus primarily in Linked Open Data. The framework covers only specific aspects, e.g. format or encoding used to publish the data and do not cover other aspects related with data values itself. The issue of badges provided by the Open Data Institute is another method for assess data quality, the main focus are metadata aspects instead of the data. In the literature, the proposed frameworks by some authors are mainly oriented to data producers.

The principal technological limitation (RQ3) identified is related with the heterogeneity of the open datasets due to different data sources, data types and the difficulty associated with their integration. The heterogeneous data formats and standards difficult interoperability, which makes quality assessments more complex.

According to the literature review, tools used for data quality assessment (RQ4) can be grouped as commercial and non-commercial. There are 16 commercial tools examined by Gartner and classified in the Magic Quadrant for Data Quality Solutions. These tools require a paid subscription for usage. Great Expectations is an open-source library use for data assessment, facilitating the validation of data quality in accordance with predetermined expectations. This tool was choosen as data quality tool to be integrated into proposed framework.

The aim of this dissertation therefore is to propose a framework for evaluating the quality of open data across four data quality dimensions. In the first part, systemic and technical aspects of the open data environment (Accessibility, Interoperability) and in the second part, two inherent data quality dimensions (Completeness, Uniqueness) of

the datasets were assessed (RQ5).

A set of visualizations were developed for presentation of the results obtained in a range from 0 to 100 points. Figure 5.1 presents the behavior of the overall quality indicator (77 points out of 100 possible points). It is evident that the indicator show data quality problems in its value is under the Portugal Open Data Quality Score (89).

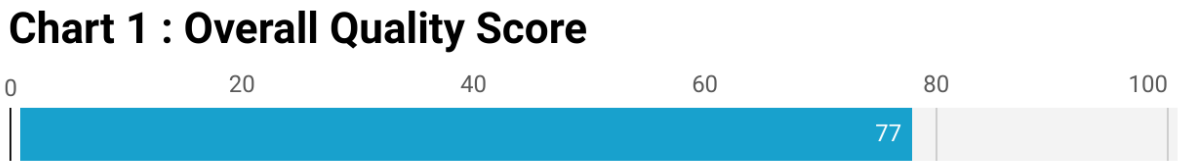


Figure 5.1. Overall Data Quality Score
Source : By the author

Results of the assessment shown that overall quality 6 datasets out of 8 assessed are under Portugal Open Data Quality Score (89). Two datasets present a high level of quality (95.50), both datasets are published under the health domain (SNS) as depicted in Figure 5.2.

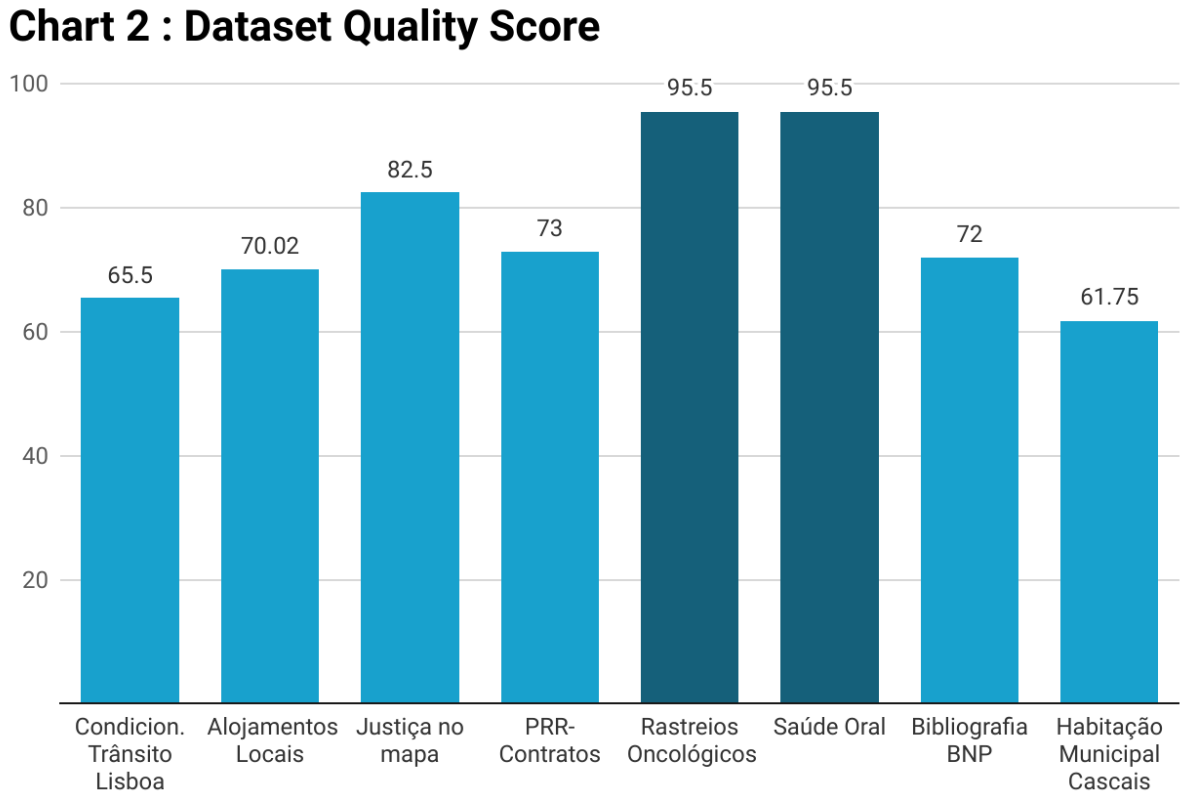


Figure 5.2. Dataset Data Quality Score
Source : By the author

In order to comprehend this behavior, Figure 5.3 illustrates the computed dimensions along with an enhanced examination of these outcomes, demonstrating that the

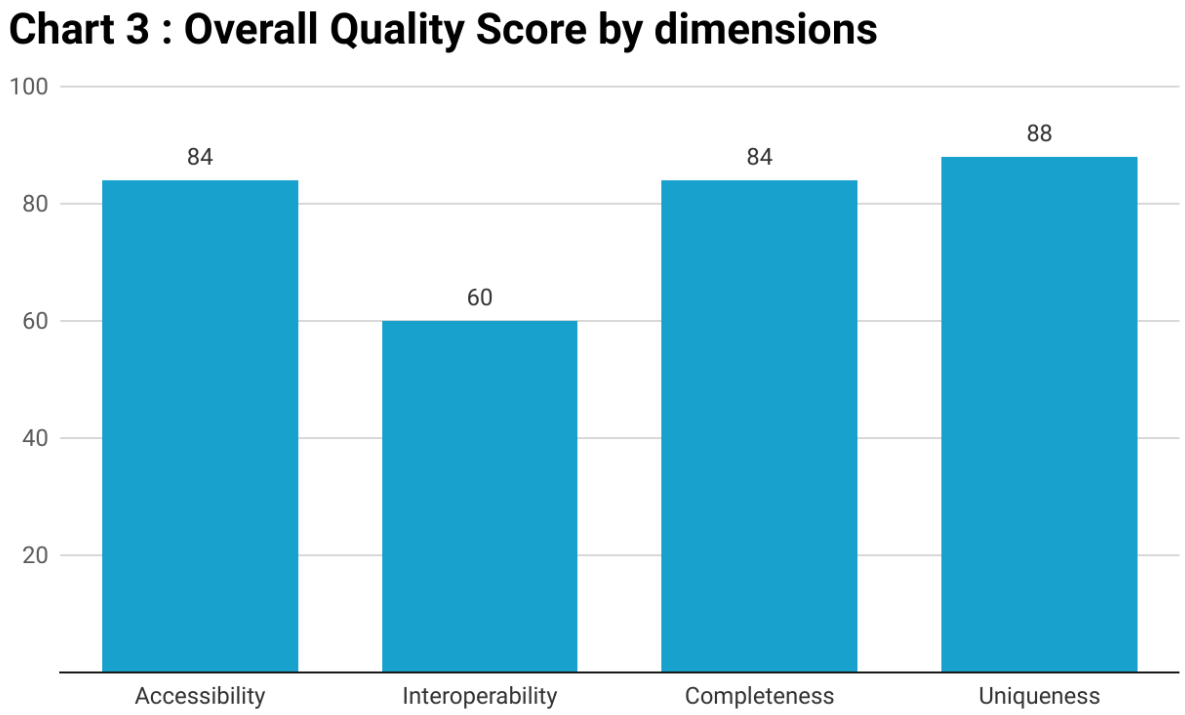


Figure 5.3. Overall Quality Score by dimension
Source : By the author

dimensions which encountered issues are as presented in Table 5.1.

Some quality problems were detected impacting in the final quality score of the datasets as outlined in the Table 5.1.

Table 5.1. Portuguese Open Datasets Assessment Results

		Condic. Trânsito	Alojamentos	Justiça no mapa	PRR-Contratos	Rastreios Onc.	Saúde Oral	Bibliografia	Habitação Mun.
Quality Dimension	**Points**	**Points Scored**	**Points Scored**	**Points Scored**	**Points Scored**	**Points Scored**	**Points Scored**	**Points Scored**	**Points Scored**
Accessibility	30	24	21	24	24	30	30	24	24
Interoperability	30	16.50	15	27	9	25.5	25.5	9	15
Completeness	25	25	19.02	16.50	25	25	25	24	7.75
Uniqueness	15	0.0045	15	15	15	15	15	15	15
Total Quality Score		**65.50**	**70.02**	**82.50**	**73.00**	**95.50**	**95.50**	**72.00**	**61.75**

The Accessibility dimension performed good (Figure 5.3). At level of it metrics, an average of 25 points out of 30 possible points was computed, Accessible for download is the lowest scored as depicted in Figure 5.4, datasets are easily manually downloadable without complex procedures. For non-IT experts, manually download work as expected. Only 2 datasets are able to be accessed by applications via API limiting consuming of dataset by software agents. One critical issue in this dimension identified in one dataset is the access to aggregate file instead of individual files (year), this is time-consuming and impact performance, sometimes, the file is not able to download due to timeout.

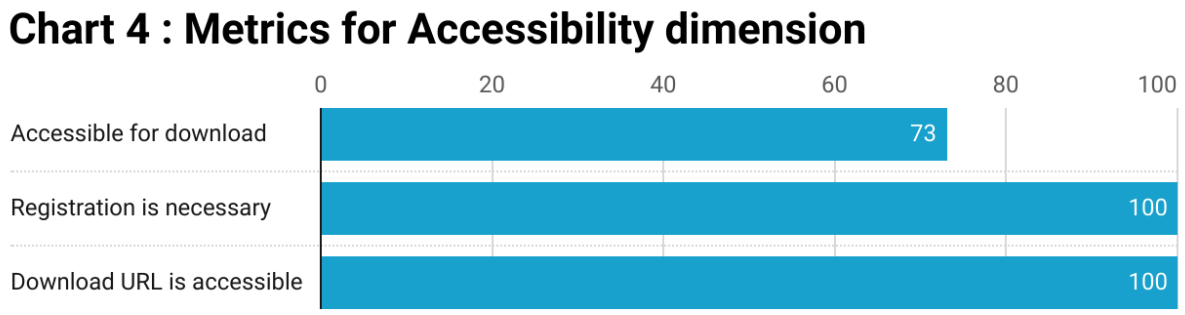


Figure 5.4. Metrics for Accessibility dimension
Source : By the author

The possibility to understand and reuse the data is a characteristic of the Interoperability. This dimension is the where the open datasets assessed performed worst (Figure 5.3). The average of points was 18 out of 30. The Interoperability Maturity metric of this dimension is the lowest as depicted in Figure 5.5. A number of 5 out of 8 datasets are schema-less or provided with basic definition of fields, therefore, users are not able to interpret and reuse data correctly. Two datasets were accompanied by field names and data-schema in JSON. Three datasets evaluated do not included meta-data documentation that described the fields of the dataset. The user has difficulties in understanding the data in few datasets because there are field names with acronyms, uncommon or technical words. Additionally, it was observed that open data infrastructure commonly do not provide contextual information pertaining to the datasets that are provided.

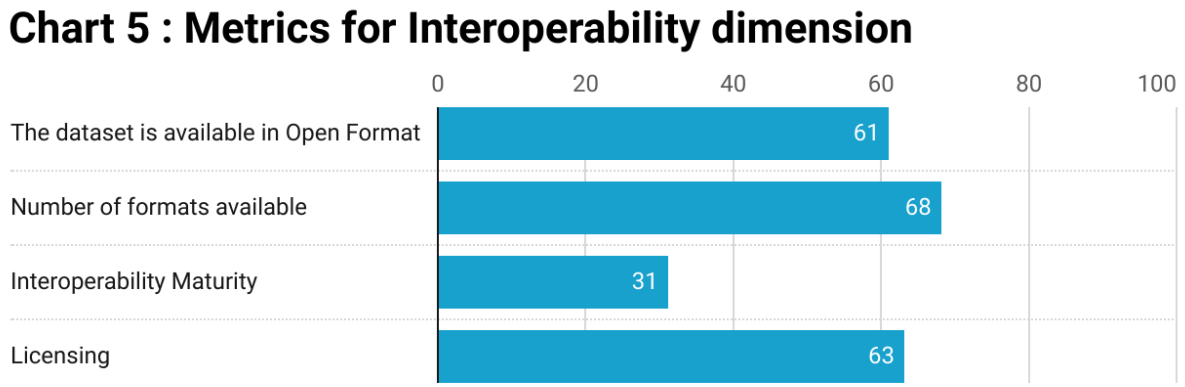


Figure 5.5. Metrics for Interoperability dimension
Source : By the author

The Completeness dimension exhibited good performance (Figure 5.3), this dimension only have one metric as depicted in Figure 5.6. The score achieved in average was

21 points from a possible 25. There exist two explanations: either the data published is either consolidated or statistical, as exhibited in the datasets "Rastreios Oncológicos" or "Saúde Oral," or this is attributable to the nature of the published data, which contains field names that can be readily inferred the type of data and the domain, thereby allowing for the selection of columns to be evaluated according to the proposed framework.

Figure 5.6. Metric for Completeness dimension
Source : By the author

The dimension Uniqueness is the one where the assessment the open dataset performed best (Figure 5.3), this dimension only have one metric as depicted in Figure 5.7 and was scored with 13 out of 15 points. Since all datasets available from the open data portals consisted of a single table (provided for download as CSV, XSLX, XML, JSON files), hence the dimension of Uniqueness based on key field or self-increasing key was considered in this evaluation by Great Expectations in order to identify that the data be uniquely identifiable. Datasets without key field or row ID identification, the datasets were assessed using compound columns collectively.

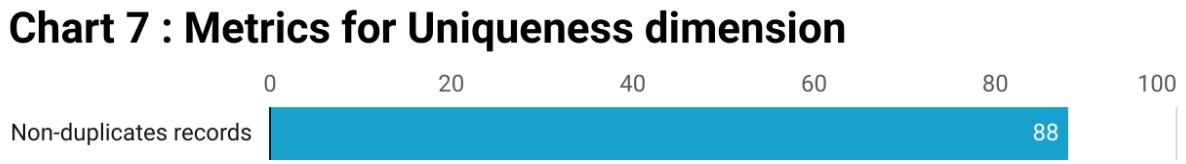


Figure 5.7. Metric for Uniqueness dimension
Source : By the author

CHAPTER 6

Conclusions

Governments across the globe are encouraging public entities to publish their data for public access. The prospects of Open Data are exceedingly significant. Nevertheless, despite the initiation of the Open Data movement several years prior, there remain certain challenges that must still be addressed. Data quality issues are a barrier to adopt Open Data.

This dissertation proposes a framework for assess open data quality based on representative dimensions of inherent data quality and aspects related to open data infrastructure identified in the systematic review of papers. We evaluated the proposed framework by applying it on OGD dataset from different domains published in Portuguese Open Data Portals.

The majority of the datasets (80%) do not allow download using API, this issue can limit the adopt open data in user's applications due to limited reusing due to many users prefer use of API to include the data directly into their application.

We observed common data quality issues impacting the Interoperability. The majority of the datasets do not contains minimum documentation describing the attributes in the data that was provided with the dataset. The data was accompanied only by field names and without metadata documentation. Exist significant heterogeneity between the datasets in terms of character encoding and date formats, which had an impact on the automatic calculation of the Uniqueness and Completeness dimension. For certain datasets, manual intervention was required. This heterogeneity indicates absence to adherence to commons standards. There exists a lack of a universally accepted standard for the publication of open datasets. This reality serves to complicate the reuse of open data. Certain concerns pertaining to data quality continue to persist. Moreover, although access to open datasets may be feasible, it does not necessarily imply that the information is suitable for reuse.

The framework proposed has been developed as an solution to assess open data quality. It provides a quality score based on dimensions and its metrics. An open-source quality tool is an important component of the framework permitting asses inherent data quality. Currently the proposed framework is only able to assess datasets published in tabular formats. It could be improved in order to support another formats.

One of the limitations of this dissertation was that many of the articles reviewed were orientated towards Linked Open Data. Other limitations, for example, related to the datasets analysed, required additional work to resolve technical problems related

to encoding and field delimiters prior to their evaluation by the framework. The researcher's inexperience using Great Expectations, which requires additional effort for its use in the proposed framework, was another limitation we faced.

6.1. Future Work

Looking to the future, opportunities arise to improve the proposed framework. Making the framework also applicable for additional intrinsic quality dimension like Consistency. For instance the dimensions and relevant metrics chosen for the framework are not able to detect valid values in columns with public domains.

Additionally, we intend to broaden the application of Great Expectations (GE) within the framework to encompass additional data domains pertinent to Open Government Data. This quality tool facilitates the establishment of tailored expectations in accordance with criteria that cannot be evaluated utilizing native expectations.

Finally, the proposed framework could be complemented expanding the framework's applicability to non-tabular data .

References

[1] Abdulrazzak Ali, Nurul Akmar Emran, Siti Azirah Asmai, and Amelia Ritahani Ismail. "An Assessment of Open Data Sets Completeness". In: *International Journal of Advanced Computer Science and Applications* (2019).

[2] Morgana Carneiro Andrade, Rafaela Oliveira da Cunha, Jorge Figueiredo, and Ana Alice Baptista. "Do the european data portal datasets in the categories government and public sector, transport, and education, culture and sport meet the data on the web best practices?" In: *Data* 6.8 (2021).

[3] Sören Auer. "Introduction to LOD2". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8661 (2014), pp. 1–17.

[4] C. Batini and M. Scannapieco. "Data Quality Issues in Linked Open Data". In: *Data and Information Quality*. Data-Centric Systems and Applications. Springer International Publishing AG Switzerland, 2016, pp. 87–112. ISBN: 978-3-319-24104-3.

[5] Behshid Behkamal, Mohsen Kahani, Ebrahim Bagheri, and Zoran Jeremic. "A metrics-driven approach for quality assessment of linked open data". In: *Journal of Theoretical and Applied Electronic Commerce Research* 9.2 (2014), pp. 64–79.

[6] Martin Beno, Kathrin Figl, Jürgen Umbrich, and Axel Polleres. "Perception of key barriers in using and publishing open data". In: *eJournal of eDemocracy and Open Government* 9.2 (2017), pp. 134–165.

[7] Tim Berners-Lee. *Design issues: Linked data*. URL: `https://www.w3.org/DesignIssues/LinkedData.html`. (accessed: 26.01.2024).

[8] Francisco Javier Campos Zabala. "How to Approach Open Data". In: *Grow Your Business with AI: A First Principles Approach for Scaling Artificial Intelligence in the Enterprise*. Berkeley, CA: Apress, 2023, pp. 307–325. ISBN: 978-1-4842-9669-1.

[9] Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. "Evaluating the quality of linked open data in digital libraries". In: *Journal of Information Science* 48.1 (2022), pp. 21–43.

[10] Paulo Da Silva Carvalho, Patrik Hitzelberger, Fatma Bouali, and Gilles Venturini. "A Visual Technique to Assess the Quality of Datasets-Understanding the Structure and Detecting Errors and Missing Values in Open Data CSV Files". In: *International Conference on Data Management Technologies and Applications*. Vol. 2. SCITEPRESS. 2015, pp. 134–141.

[11] Michael Chiu, Diana Farrel, Peter Groves, James Manyika, Peter Groves, Steve Van Kuiken, and Elizabeth Almasi Doshi. *Open data: Unlocking innovation and performance with liquid Information*. Oct. 2013. URL: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20information/mgi_open_data_fullreport_oct2013.pdf. (accessed: 30.03.2024).

[12] Michael Chiu, Diana Farrel, and Kate Jackson. *How government can promote open data and help unleash over $3 trillion in economic values*. Apr. 2014. URL: https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/How%20government%20can%20promote%20open%20data/How%20government%20can%20promote%20open%20data.pdf. (accessed: 30.03.2024).

[13] Adrienne Colborne and Michael Smit. "Identifying and mitigating risks to the quality of open data in the post-truth era". In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 2588–2594.

[14] David Corsar and Peter Edwards. "Challenges of open data quality: More than just license, format, and customer support". In: *Journal of Data and Information Quality* 9.1 (2017).

[15] Thomas H. Davenport and Thomas C. Redman. "How AI Is Improving Data Management". English. In: *MIT Sloan Management Review* 64.2 (2023), pp. 1–5.

[16] Renato De Donato, Giuseppe Ferretti, Antonio Marciano, Giuseppina Palmieri, Donato Pirozzi, Vittorio Scarano, and Luca Vicidomini. "Agile production of high quality open data". In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. Delft, The Netherlands: Association for Computing Machinery, 2018. ISBN: 9781450365260.

[17] Laurence Dierickx. "News bot for the newsroom: how building data quality indicators can support journalistic projects relying on real-time open data". In: *Global Investigative Journalism Conference 2017 Academic Track*. 2017.

[18] Publications Office European Union. *Data.europa.eu data quality guidelines*. Publications Office, 2021.

[19] Publications Office of the European Union, M Assen, G Cecconi, G Carsaniga, E Lincklaen Arriëns, and J Dogger. *Open data maturity report 2023*. Publications Office of the European Union, 2023.

[20] Aurora González-Vidal, Alfonso P. Ramallo-González, and Antonio F. Skarmeta. "Intrinsic and extrinsic quality of data for open data repositories". In: *ICT Express* 8.3 (2022), pp. 328–333.

[21] Jose Gurin. *Big data and open data: what's what and why does it matter?* URL: https://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government. (accessed: 10.01.2024).

[22] Anne Immonen, Eila Ovaska, and Tuomas Paaso. "Towards certified open data in digital service ecosystems". In: *Software Quality Journal* 26.4 (2018), pp. 1257–1297.

[23] ISO. *ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. en. Standard. URL: https://www.iso.org/standard/35736.html.

[24] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. "Benefits, Adoption Barriers and Myths of Open Data and Open Government". In: *Information Systems Management* 29.4 (2012), pp. 258–268.

[25] Matti Keränen et al. "Knowledge processes and information quality in open data context: conceptual considerations and empirical findings". Master's Thesis. LAPPEENRANTA UNIVERSITY OF TECHNOLOGY, 2017.

[26] Rob Kitchin. *The Data Revolution - Big Data,Open Data, Data Infrastructures and their consequences*. SAGE Publications, 2014. ISBN: 978-1-4462-8747-7.

[27] Tibor Koltay. "Quality of open research data: Values, convergences and governance". In: *Information (Switzerland)* 11.4 (2020).

[28] Pavel Krasikov and Christine Legner. "A Method to Screen, Assess, and Prepare Open Data for Use". In: *Journal of Data and Information Quality* 15.4 (2023).

[29] Sylvain Kubler, Jérémy Robert, Yves Le Traon, Jörgen Umbrich, and Sebastian Neumaier. "Open data portal quality comparison using AHP". In: vol. 08-10-June-2016. 2016, pp. 397–407.

[30] Sylvain Kubler, Jérérmy Robert, Sebastian Neumaier, Jürgen Umbrich, and Yves Le Traon. "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process". In: *Government Information Quarterly* 35.1 (2018). Internet Plus Government: Advancement of Networking Technology and Evolution of the Public Sector, pp. 13–29. ISSN: 0740-624X.

[31] Dasa Kusnirakova, Mouzhi Ge, Leonard Walletzky, and Barbora Buhnova. "Interoperability-oriented Quality Assessment for Czech Open Data". In: *Proceedings of the 11th International Conference on Data Science, Technology and Applications*. SCITEPRESS - Science and Technology Publications, 2022.

[32] Sara Laurita, Salvatore Iiritano, and Mariagrazia Zottoli. "Data Quality Issue: The Open Data Explorer Solution". In: *Conference for E-Democracy and Open Government*. 2016, pp. 33–38.

[33] Georg J. P. Link et al. "Contemporary issues of open data in information systems research: Considerations and recommendations". In: *Communications of the Association for Information Systems* 41.1 (2017), pp. 587–610.

[34] David Loshin. *Enterprise Knowledge Management - The Data Quality Approach*. Morgan Kaufmann Publishers, 2003. ISBN: 0-12-455840-2.

[35] Rupa Mahanti. "Data quality and data quality dimensions". In: *Software Quality Professional* 22.1 (2019), pp. 4–8.

[36] André Drumond Mateus. "Public Policy: Turning Open Data into Democratic Data-Portal Quality Assessment-Comparative Analysis". Master's Thesis. Universidade NOVA de Lisboa (Portugal), 2023.

[37] Laura María Gutiérrez Medina, José Luis Roda García, Guillermo García Juanes, Alioth Rodríguez Barrios, and Pedro González Yanes. "Open data strategies and experiences to improve sharing and publication of public sector information". In: *eJournal of eDemocracy and Open Government* 6.1 (2014), pp. 80–86.

[38] Microsoft. *Introduction to Data Quality Services*. URL: https://learn.microsoft.com/en-us/sql/data-quality-services/introduction-to-data-quality-services?view=sql-server-ver16. (accessed: 30.03.2024).

[39] Maryam Moghadami and Mila Malekolkalami. "Evaluating the quality of open data portals in Iran". In: *International journal of knowledge processing studies* 2.1 (2022), pp. 43–53.

[40] Jean-Louis Monino and Soraya Sedkaoui. *Big Data, Open Data and Data Development*. Smart Innovation Set volume 3. Wiley-ISTE, 2016. ISBN: 978-1-84821-880-2.

[41] Domenico Natale. "Extensions of ISO/IEC 25000 Quality Models to the Context of Artificial Intelligence". In: *Proceedings of IWESQ@ APSEC* (2022).

[42] Sebastian Neumaier. "Open data quality: Assessment and evolution of (meta-) data quality in the open data landscape". PhD thesis. Wien, 2015.

[43] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. "Automated quality assessment of metadata across open data portals". In: *Journal of Data and Information Quality* 8.1 (2016).

[44] Anastasija Nikiforova. "Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment". In: *Baltic Journal of Modern Computing* 8.3 (2020), pp. 391–432.

[45] Anastasija Nikiforova. "Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia". In: *Baltic Journal of Modern Computing* 6.4 (2018). ISSN: 2255-8950.

[46] Anastasija Nikiforova, Zane Bicevska, Janis Bicevskis, and Ivo Oditis. "Data quality evaluation: a comparative analysis of company registers' open data in four European countries". In: Sept. 2018, pp. 197–204.

[47] Anastasija Nikiforova and Janis Bicevskis. "Open data quality". In: vol. 2158. 2018, pp. 151–160.

[48] Anastasija Nikiforova, Janis Bicevskis, Zane Bicevska, and Ivo Oditis. "User-oriented approach to data quality evaluation". In: *Journal of Universal Computer Science* 26.1 (2020), pp. 107–126.

[49] Jack E. Olson. *Data Quality - The Accuracy Dimension*. Morgan Kaufmann Publishers, 2003. ISBN: 1-55860-891-5.

[50] OpenDataHandbook. *What is Open Data?* URL: https://opendatahandbook.org/guide/en/what-is-open-data/. (accessed: 15.12.2023).

[51] OpenDefinition. *Open Definition 2.1*. 2024. URL: https://opendefinition.org/od/2.1/en/. (accessed: 12.01.2024).

[52] OpenKnowledge. *The Open Data Handbook*. URL: https://opendatahandbook.org/. (accessed: 12.01.2024).

[53] Edgar Oviedo, Jose Norberto Mazon, and Jose Jacobo Zubcoff. "Towards a data quality model for open data portals". In: *Proceedings of the 2013 39th Latin American Computing Conference, CLEI 2013* (2013).

[54] Governo de Portugal. *Portal de dados abertos da Administração Pública*. URL: https://dados.gov.pt/pt/datasets/. (accessed: 30.07.2024).

[55] Vigan Raca, Goran Velinov, Betim Cico, and Margita Kon-Popovska. "Application-based Framework for Analysis, Monitoring and Evaluation of National Open Data Portals". In: *International Journal of Advanced Computer Science and Applications* 12.11 (2021), pp. 26–36.

[56] MOHAMMAD RIFAT AHMMAD Rashid, Marco Torchiano, et al. "A systematic literature review of open data quality in practice". In: *Open Data Research Symposium*. Open Data Research Symposium. 2016.

[57] Guang-Jie Ren and Susanne Glissmann. "Identifying information assets for open data: the role of business architecture and information quality". In: *2012 IEEE 14th International Conference on Commerce and Enterprise Computing*. IEEE. 2012, pp. 94–100.

[58] Shazia Sadiq and Marta Indulska. "Open data: Quality over quantity". In: *International Journal of Information Management* 37.3 (2017), pp. 150–154.

[59] Cesar Garcia Saez. "Improving open data quality through citizen engagement and data engineering". In: *ACM International Conference Proceeding Series*. 2022.

[60] Ina Schieferdecker. "(Open) Data Quality". In: *2012 IEEE 36th Annual Computer Software and Applications Conference*. 2012, pp. 83–84.

[61] Laura Sebastian-Coleman. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Elsevier, 2013. ISBN: 978-0-12-397033-6.

[62] Laura Sebastian-Coleman. *Meeting the Challenges of Data Quality Management*. Elsevier Science and Technology Books, Inc., 2022. ISBN: 978-0-12-821737-5.

[63] Sudesh Sheoran, Saravanan Mohanasundaram, R. Kasilingam, and S.K. Vij. "Usability and Accessibility of Open Government Data Portals of Countries Worldwide: An Application of TOPSIS and Entropy Weight Method". In: *International Journal of Electronic Government Research* (2023).

[64] Barbara Slibar, Dijana Oreski, and Bozidar Klicek. "Aspects of open data and illustrative quality metrics: literature review". In: *Economic and Social Development: Book of Proceedings* (2018), pp. 90–99.

[65] Barbara Šlibar, Dijana Oreški, and Nina Begičević Ređep. "Importance of the Open Data Assessment: An Insight Into the (Meta) Data Quality Dimensions". In: *SAGE Open* 11.2 (2021).

[66] Chang-Suk Song, Go-Eun Park, and Chang-Jae Kim. "Quality Characteristics of Public Open Data". In: *Advanced Science and Technology Letters* 139 (2016), pp. 103–107.

[67] Prashanth H. Southekal. *Data Quality—Empowering Businesses with Analytics and AI*. John Wiley and Sons (US), 2023. ISBN: 978-1-39416-523-0.

[68] Milena Stróżyna, Gerd Eiden, Witold Abramowicz, Dominik Filipiak, Jacek Małyszko, and Krzysztof Węcel. "A framework for the quality-based selection and retrieval of open data - a use case from the maritime domain". In: *Electronic Markets* 28.2 (2018), pp. 219–233.

[69] John K. Thompson. "Moving From Open Data to Our Data". In: *Data for All*. Manning, 2023. ISBN: 978-1633438774.

[70] Open Data Toronto. *Towards an updated Data Quality Score in Open Data*. Aug. 2023. URL: `https://open.toronto.ca/towards-an-updated-data-quality-score-in-open-data/`. (accessed: 13.03.2024).

[71] Ana Trisovic, Katherine Mika, Ceilyn Boyd, Sebastian Feger, and Mercè Crosas. "Repository approaches to improving quality of shared data and code". In: *Data* 6.2 (2021), pp. 1–12.

[72] Sander Van Der Waal, Krzysztof W Ę Cel, Ivan Ermilov, Valentina Janev, Uro Š Milošević, and Mark Wainwright. "Lifting open data portals to the data web". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8661 (2014), pp. 175–195.

[73] Antonio Vetrò, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, and Federico Morando. "Open data quality measurement framework: Definition and application to Open Government Data". In: *Government Information Quarterly* 33.2 (2016), pp. 325–337.

[74] Lisa Wenige, Claus Stadler, Michael Martin, Richard Figura, Robert Sauter, and Christopher W Frank. "Open Data and the Status Quo–A Fine-Grained Evaluation Framework for Open Data Quality and an Analysis of Open Data Portals in Germany". In: *arXiv preprint arXiv:2106.09590* (2021).

[75] Ruojing Zhang, Marta Indulska, and Shazia Sadiq. "Discovering Data Quality Problems: The Case of Repurposed Data". In: *Business and Information Systems Engineering* 61.5 (2019), pp. 575–593.

[76] Anneke Zuiderwijk, Marijn Janssen, and Iryna Susha. "Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators". In: *Journal of Organizational Computing and Electronic Commerce* 26.1-2 (2016), pp. 116–146.

APPENDIX A

Studies analysed in the Literature Review.

Table A.1: Studies analysed in the Literature Review.

Id	Ref.	Year	Author(s)	Title	Type
1	[60]	2012	Ina Schieferdecker	(Open) Data Quality	Conference Proceedings
2	[5]	2014	Behkamal et al.	A metrics-driven approach for quality assessment of linked open data	Journal Article
3	[68]	2018	Stróżyna et al.	A framework for the quality-based selection and retrieval of open data-a use case from the maritime domain	Journal Article
4	[28]	2019	Krasikov & Legner	A Method to Screen, Assess, and Prepare Open Data for Use	Journal Article
5	[56]	2016	Torchiano, et al.	A systematic literature review of open data quality in practice	Conference Proceedings
6	[10]	2015	Carvalho et al.	A Visual Technique to Assess the Quality of Datasets-Understanding the Structure and Detecting Errors and Missing Values in Open Data CSV Files	Conference Proceedings
7	[16]	2018	De Donato et al.	Agile production of high quality open data	Conference Proceedings
8	[1]	2019	Ali et al.	An Assessment of Open Data Sets Completeness	Journal Article
9	[55]	2021	Raca et al.	Application-based Framework for Analysis, Monitoring and Evaluation of National Open Data Portals	Journal Article
10	[64]	2018	Slibar et al.	Aspects of open data and illustrative quality metrics: literature review	Journal Article
11	[43]	2016	Neumaier et al.	Automated quality assessment of metadata across open data portals	Journal Article
12	[24]	2012	Janssen et al.	Benefits, Adoption Barriers and Myths of Open Data and Open Government	Journal Article
13	[14]	2017	Corsar & Edwards	Challenges of Open Data Quality: More Than Just License, Format, and Customer Support	Journal Article
14	[30]	2018	Kubler et al.	Comparison of metadata quality in open data portals using the Analytic Hierarchy Process	Journal Article
15	[33]	2017	Georg J. P. Link et al.	Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations	Journal Article
16	[35]	2019	Rupa Mahanti	Data Quality and Data Quality Dimensions	Journal Article
17	[46]	2018	Nikiforova et al.	Data quality evaluation: a comparative analysis of company registers' open data in four European countries.	Journal Article
18	[32]	2018	Laurita et al.	Data Quality Issue: The Open Data Explorer Solution	Conference Proceedings

Continued on next page

Table A.1: Studies analysed in the Literature Review. (Continued)

Id	Ref.	Year	Author(s)	Title	Type
19	[44]	2020	Anastasija Nikiforova	Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment	Journal Article
20	[75]	2019	Zhang et al.	Discovering Data Quality Problems: The Case of Repurposed Data	Journal Article
21	[2]	2021	Andrade et al.	Do the european data portal datasets in the categories government and public sector, transport, and education, culture and sport meet the data on the web best practices?	Journal Article
22	[9]	2022	Candela et al.	Evaluating the quality of linked open data in digital libraries	Journal Article
23	[39]	2022	Moghadami & Malekolkalami	Evaluating the quality of open data portals in Iran	Journal Article
24	[41]	2022	Domenico Natale	Extensions of ISO/IEC 25000 Quality Models to the Context of Artificial Intelligence	Conference Proceedings
25	[13]	2017	Colborne & Smit	Identifying and mitigating risks to the quality of open data in the post-truth era	Conference Proceedings
26	[57]	2012	Ren & Glissmann	Identifying Information Assets for Open Data - The Role of Business Architecture and Information Quality	Conference Proceedings
27	[15]	2023	Davenport & Redman	How AI Is Improving Data Management	Journal Article
28	[12]	2023	Chiu et al.	How government can promote open data and help unleash over $3 trillion in economic values	Journal Article
29	[65]	2021	Šlibar et al.	Importance of the Open Data Assessment: An Insight Into the (Meta) Data Quality Dimensions	Journal Article
30	[76]	2016	Zuiderwijk et al.	Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators	Journal Article
31	[59]	2022	Cesar Garcia Sae	Improving open data quality through citizen engagement and data engineering	Conference Proceedings
32	[31]	2023	Kusnirakova et al.	Interoperability-oriented Quality Assessment for Czech Open Data	Journal Article
33	[20]	2022	González-Vidal et al.	Intrinsic and extrinsic quality of data for open data repositories	Journal Article
34	[3]	2014	Sören Auer	Introduction to LOD2	Journal Article
35	[25]	2019	Matti Keränen	Knowledge processes and information quality in open data context: conceptual considerations and empirical findings	Master Thesis
36	[72]	2014	Van Der Waal et al.	Lifting open data portals to the data web".	Journal Article
37	[17]	2017	Laurence Dierickx	Research: "News bot for the newsroom: how building data quality indicators can support journalistic projects relying on realtime open data"	Conference Proceedings
38	[19]	2023	Publications Office of the European Union	Open data maturity report 2023	Report
39	[47]	2018	Nikiforova & Bicevskis	Open data quality	Conference Proceedings
40	[73]	2016	Vetrò et al.	Open data quality measurement framework: Definition and application to Open Government Data	Journal Article

Continued on next page

Table A.1: Studies analysed in the Literature Review. (Continued)

Id	Ref.	Year	Author(s)	Title	Type
41	[45]	2020	Anastasija Nikiforova	Open Data Quality Evaluation: A comparative analysis of open data in Latvia	Journal Article
42	[42]	2015	Sebastien Neumaier	Open data quality: assessment and evolution of (meta-) data quality in the open data landscape	Master Thesis
43	[37]	2014	Medina et al.	Open data strategies and experiences to improve sharing and publication of public sector information	Journal Article
44	[29]	2016	Kubler et al.	Open data portal quality comparison using AHP	Conference Proceedings
45	[74]	2021	Wenige et al.	Open Data and the Status Quo–A Fine-Grained Evaluation Framework for Open Data Quality and an Analysis of Open Data Portals in GermanyAHP	Journal Article
46	[58]	2017	Sadiq & Indulska	Open data: Quality over quantity	Journal Article
47	[6]	2017	Beno et al.	Perception of key barriers in using and publishing open data	Journal Article
48	[36]	2023	André Drumond Mateus	Public Policy: Turning Open Data into Democratic Data-Portal Quality Assessment-Comparative Analysis	Master Thesis
49	[66]	2016	Song et al.	Quality Characteristics of Public Open Data	Journal Article
50	[27]	2020	Tibor Koltay	Quality of Open Research Data: Values, Convergences and Governance	Journal Article
51	[71]	2021	Trisovic et al.	Repository approaches to improving quality of shared data and code	Journal Article
52	[53]	2013	Oviedo et al.	Towards a data quality model for open data portals	Conference Proceedings
53	[22]	2018	Immonen et al.	Towards certified open data in digital service ecosystems	Journal Article
54	[63]	2023	Sheoran et al.	Usability and Accessibility of Open Government Data Portals of Countries Worldwide: An Application of TOPSIS and Entropy Weight Method	Journal Article
55	[48]	2020	Nikiforova et al.	User-oriented approach to data quality evaluation	Journal Article