IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 May 2023, accepted 6 June 2023, date of publication 13 June 2023, date of current version 22 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285611

RESEARCH ARTICLE

An Assessment of the Quality of Open Government Data in Saudi Arabia

NADA FAISAL ALOGAIEL[1] AND OMER ABDULAZIZ ALRWAIS[2]
[1]Department of Information Systems, College of Computer and Information Systems, Imam Mohammad Ibn Saud Islamic University, Riyadh 11564, Saudi Arabia
[2]Department of Information Systems, College of Computer and Information Systems, King Saudi University, Riyadh 11495, Saudi Arabia

Corresponding author: Nada Faisal Alogaiel (nfalogaiel@imamu.edu.sa)

ABSTRACT Open government data (OGD) is an e-governance practice that aims to increase public resource efficiency and improve service delivery for citizens. Saudi Arabia launched its first open data initiative in 2011, with the goal of maximizing the economic impact of open data locally. This empirical study aimed to assess the effects of the Saudi OGD initiative by conducting a literature review and monitoring the impact of Saudi state-of-the-art open data. After observation and analysis, a systematic Monitor, Analysis, Action, and Review (MAAR) approach was followed to conduct a proposed solution that addresses quality shortcomings that consequently affect open data reusability. The methodology for creating the proposed solution included selecting, formulating, and weighing the quality characteristics that define a valuable OGD. The proposed solution is a framework for assessing the OGD quality using metrics that are compatible with the Saudi Open Data Portal (od.data.gov.sa). To review this proposed quality assessment framework and test whether it achieved its predicted outcomes, the results of the proposed solution were evaluated using local data samples from the Saudi Open Data Portal.

INDEX TERMS E-government, framework, measurement, OGD, open data, open government data, portal, quality, Saudi Arabia.

I. INTRODUCTION

Saudi Arabia is witnessing a substantial shift towards digitalizing its government infrastructure as part of its 2030 vision plan to enhance its effectiveness. This includes adopting an open government data (OGD) approach, which is believed to enhance the transparency and accountability between the government sector and individuals [1]. Granting data transparency is more than that of public service. It is an industry with a direct market size that reached 184.45 billion EUR in 2019 for countries in the European Union and is expected to reach 199.51-334.20 billion EUR by 2025 [2]. The European experience demonstrates the potential of effectively utilizing open data for a sustainable infrastructure that enables the economy instead of a project that generates rapid revenue. However, obtaining incremental benefits requires continuous examinations and maintenance.

This study follows an empirical structure to follow the problem-solving process by Kolodner et al. [3], It starts with interpreting the problem of the OGD in Saudi Arabia by examining the literature review, followed by generating a solution in the methodology to the leading cause of the OGD shortfall, which is linked to data quality. Subsequently, the results of the proposed solution are evaluated to determine if the proposed quality assessment framework achieves the predicted outcomes, which accurately investigates the quality of the Saudi OGD.

Therefore, this study aims to investigate Saudi Arabia's OGD initiative by addressing the following questions:

RQ1. What impact has the Saudi open government data portal made so far?
RQ2. How can a portal's performance be improved to assist it in achieving its objectives?
RQ3. What benefits can come from investment in the open data infrastructure?

II. LITERATURE REVIEW

This literature review covers four main sections. It starts with the historical background of the OGD's origins and how it

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

became an international phenomenon. An analysis of the impact of the Saudi open data portal was made during its attempt to stay up with this global trend. Furthermore, the study analyzes the reasons behind the success or failure of OGD initiatives by linking open data quality and its impact. The findings highlight the need for an effective quality assessment framework to monitor OGD results. The last section covers the available quality assessment frameworks to help find a framework compatible with the needs of the Saudi Open Data Portal.

A. HISTORICAL BACKGROUND

Laws supporting opening government data can be traced back to 1766, when Sweden announced its Freedom of Press Act, which legalized accessing and publishing documents drawn by government agencies. The Act grants freedom of the press as a constitutional right, including allowing the release of official records to the public. In 1942, Merton [4] called for all scientific research to be made accessible without intellectual property restrictions. Encouraging the science and research community to adopt what is known today is an open-data approach for effective knowledge development. Another fundamental legal milestone in the OGD movement is the Freedom of Information Act (FOIA). The law was passed in the US in 1966, which gave American citizens the right to access the information records of federal agencies, unless the data were protected by law [5]. The legalization process sets a path for regulating government data publications. The term 'open data,' as used today, was first mentioned in a document published by an American scientific agency in 1995. They argued that environmental and geophysical data should be exchanged freely between countries since the weather and what affects it are essentially global phenomena, and international boundaries are thus irrelevant [6]. The data initiative reflects the influence of the computer science community. In 1998, the open-source movement drastically changed software development. People started collaborating to improve products and services and debunked the common belief that secrecy was the key to a robust system. The open-source impact inspired 30 open government advocates to gather in Sebastopol, California in 2007 to develop a set of principles for open government data. They believed in the internet's potential for data and public affairs [7]. In 2009, the former president of the US, Barak Obama, launched an open data initiative as soon as he took office [8]. After signing the Memorandum on Transparency and Open Government Act, the US government established the federal government's open data site (data.gov) in May 2009. Following that, at the United Nations General Assembly Meeting in 2011, eight countries began the Open Government Partnership (OGP) [9]. The partnership advocates transparent, participatory, inclusive, and accountable governance, inspiring other countries to join. The OGP now has 79 currently active members. Moreover, the US enacted the Digital Accountability and Transparency Act in 2014, which obligates the Department of the Treasury to standardize the reporting of financial data by other US government agencies. This law makes federal expenditures accessible and understandable to the public [10].

The history of open data reveals that such a bold idea aims to transform e-governance. Establishing an open-government state is not the end goal of this transformation; it is just a step in the ''digital government transformation'' journey, as Fig. 1 demonstrates. The movement started in the 1990s when the term ''e-governance'' emerged [11]. — The period from the late 1990s to the early 2000s witnessed the first phase of transformation. This was the era of ''e-government 1.0'', when governments started investing in information and communication technology (ICT) infrastructure. However, the traditional handling of governance operations has remained the same. The only change was the medium. The late 2000s witnessed the ''e-government 2.0'' phase, the open government era during which the public sector leaned towards collaboration and citizen participation. In the mid-2010s, efforts began toward a smart government known as ''e-government 3.0''During this time, the public sector began to utilize innovative technologies such as artificial intelligence, big data, blockchains, and the Internet of Things to enhance their services. The latest stage of digital government transformation is ''e-government 4.0'', referred to as the transformed government era. At this stage, governments can adapt to the needs of their stakeholders, regardless of whether they are citizens, businesses, or non-profit organizations [12].

The first sign of digital transformation in Saudi Arabia was observed in 2003 when a royal decree was issued, directing the Ministry of Communications and Information Technology (MCIT) to start planning to provide government services and transactions electronically. Subsequently, the Saudi government's steps to complete its e-government 1.0 transformation demonstrate that the Kingdom is inevitably moving toward the e-government 2.0 phase. The next step includes shifting the focus from an electronic government to an open government that is more citizen-oriented and encourages two-way collaboration between citizens and itself.

B. ASSESSING THE IMPACT OF THE SAUDI OPEN DATA PORTAL

The Saudi Open Data Portal (od.data.gov.sa) was established in 2014. It is one of the first data-centric platforms to form the Saudi National Data Bank (NDB), which is the backbone of the open data initiative in Saudi Arabia. The portal is one of the earliest open data projects, which makes it applicable for comprehensive evaluation. The first step in tracking its progress is to comprehend its expectations and review the portal's goals and objectives. According to the portal's open-data strategy [13], the benefits it aims to achieve are as follows:

obj1. Enhancing transparency and citizen participation.
obj2. Improving the efficiency of government services.
obj3. Providing opportunities for creating new services and products.

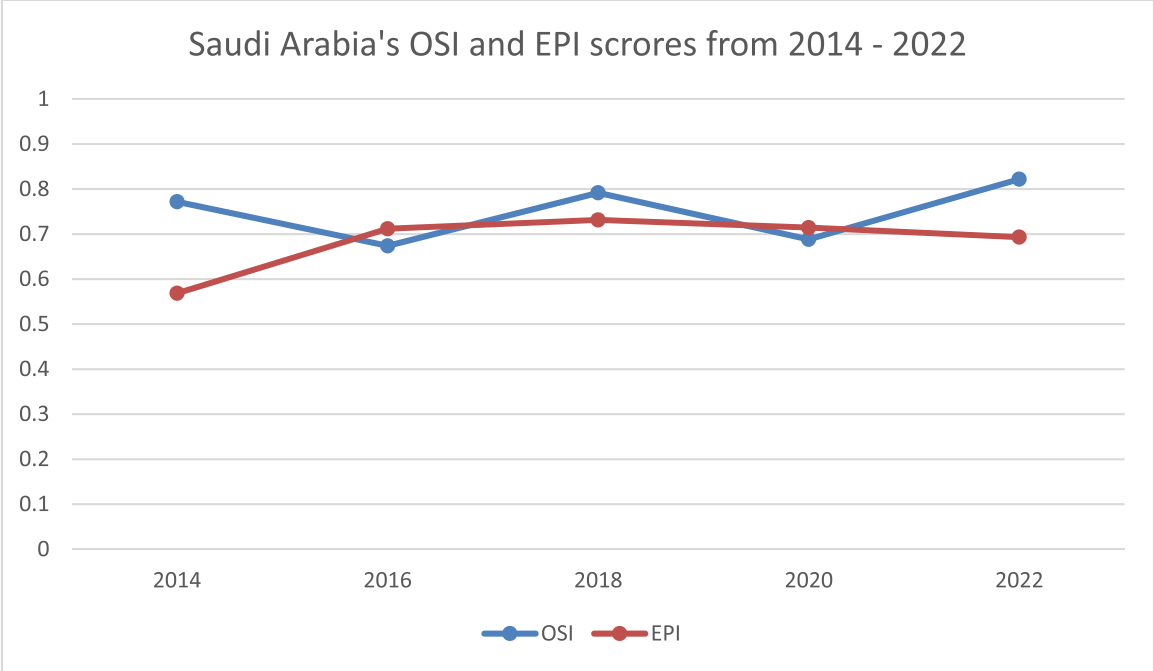

FIGURE 1. Saudi Arabia's OSI and EPI scores from 2014 – 2022.

obj4. Providing opportunities to create new jobs and economic opportunities.
obj5. Gaining new knowledge via integrating multiple data sources and processing high-volume data.

The portal's impact is evaluated by tracing the progress made so far and towards its objectives. Progress can be detected using social indicators, international benchmarking, and publication and research.

1) SOCIAL INDICATOR

The most viewed dataset had 49010 views by the time of writing this document, and the most downloaded dataset had 146 download counts. Meanwhile, 18 published use cases demonstrated the results of reusing published data. These outcomes are considered modest given that the portal has been operating for almost nine years, and 34 million Saudi residents are the main target of the platform [14].

2) INTERNATIONAL BENCHMARKING

Several OGD benchmarks track the global progress of open-data movement. Each has its own methodology and scope but shares a common goal of supporting government performance management [15]. This section tracks Saudi Arabia's ranking to assess the impact of its open-data policies.

a: OPEN DATA BAROMETER

This indicator is a universal assessment that examines the impact of open data initiatives in different countries. Each country is given a score between 0 and 100. In the latest '2018 Leaders Edition [16],' Saudi Arabia scored 25/100. Examination of the scoring methodology revealed that the score reflects quality shortcomings in open data implementation. This includes having outdated datasets, a lack of a machine-readable format to support reusability, and restricting data accessibility.

b: GLOBAL OPEN DATA INDEX

In 2015, the Open Knowledge Foundation established its last Global Open Data Index (GODI). The index ranks countries "according to their percentage of openness," [17] and Saudi Arabia ranked 103 out of 122 in the last published research. The low ranking was mainly affected by the poor open data quality, which fulfilled only 24% of the evaluation criteria.

c: OPEN DATA INVENTORY

Open Data Inventory (ODIN) rates countries' open data based on two main criteria: openness and coverage. It evaluates each country's open data compliance to international standards, and then ranks them based on their scores. In the latest report [18], Saudi Arabia ranked 64th among 193 countries. One shortcoming is the lack of administrative divisions in the data. The indicator believes that there are no data available at the second administrative level in the country, and detected gaps in the data available at the first administrative level. Moreover, the examined data are believed to lack downloading options, metadata,and associated data licenses.

d: UN E-GOVERNMENT SURVEY

The United Nations (UN) E-Government Survey considers key open data elements to assess its members' e-government status. The UN believes that inclusive people-centric analytics and applications can help with innovation and optimize resource allocation. Therefore, open data affect the

E-Participation Index (EPI) and Online Service Index (OSI) scores. Fig. 1 shows how the EPI score has been downgraded since 2016. Meanwhile, the OSI score fluctuates in the scoring system, but its ranking among other UN Members, as demonstrated in Fig. 2, has been downgraded since 2018 [19].

3) PUBLICATION AND RESEARCH

Researchers are keen to measure the effect of the global OGD movement, resulting in the development of frameworks and models to track its impact. This section reviews studies that target the Saudi open data portal. Starting with a study by Alzamil and Vasarhelyi's [20] compared Saudi Arabia and Brazil by evaluating the data transparency of their national portals. The model compares portals according to four attributes: data availability, data openness, data analytics, and application within the government website. Although neither country provided applications within the government website, Brazil is believed to have a better implementation of transparency. The transparency gap between the two countries is shown by comparing their performance in a unified GODI case, where Brazil met 80% of the evaluation criteria, whereas Saudi Arabia met only 35%.

Asyri and Al-Suraihi [21] compared open data portals across Gulf Cooperation Council (GCC) countries using a checklist model. The availability of 32 services determines the quality of a country's national open-data portals. Saudi Arabia scored 16/32 and ranked third, while the United Arab Emirates (UAE) and Oman shared first place, and Bahrain ranked second.

Saxena [22] investigated the influence of cultural differences on open data initiatives among three countries with cultural variation based on Hofstede Insights's [23] dimensions: Japan, Saudi Arabia, and the Netherlands. The study confirmed that culture does affect the reality of open data in countries. Accordingly, Saudi Arabia was more conservative with its OGD, which is reflected by its lack of encouragement for data publishing, maintenance, and reuse. In another study [24], Saxena evaluated a national OGD portal using the framework proposed by Máchová et al. [25]. They concluded that the quality of datasets is affected by the lack of visualization and mapping tools, updates on the datasets, user participation and engagement via public conversations, and data that are capable of statistical interpretation.

Finally, AlRushaid and Saudagar [26] attempted to alter the GODI Scoring Model to evaluate open data portals by adding three criteria: the use of social media, API, and existence of a mobile application. The authors evaluated the Saudi open data portal based on the altered model, compared the results with five other nations' open data portal assessments (Taiwan, United Kingdom (UK), Denmark, Colombia, and Finland), and concluded that the Saudi portal scored the lowest openness score of 32.23%.

To summarize the findings of the previous three indicators, Table 1 maps the portal's objective to the indicator that measures its progress. The analysis covered all goals except for the fourth objective, which could not be measured because of the lack of statistics tracking the economic progress of open data in Saudi Arabia. Generally, the indicators reveal the portal's mediocre performance compared with its objectives.

TABLE 1. Linking the Saudi OGD objective to the indicators used to ASSESS its accomplishment.

Indicator	Objective
2.1	obj3
2.2	obj1 , obj2
2.3	obj2, obj5

C. THE LINK BETWEEN OPEN DATA QUALITY AND ITS IMPACT

According to 'the open data impact process [27],' the origin of low impact is traced back to the publishing organizations, as shown in Fig. 3. Considering the garbage in/garbage out (GIGO) concept, an organization's poor data quality affects the impact of published data. In addition, maintaining open data quality affects sustainability not just its impact. Since low-quality data consumes resources without stimulating reusability, they will not deliver profits. Fig. 4 shows the effect cycle for low-quality data. It demonstrates how data quality affects the value of the dataset and, in turn, the return on investment (ROI). The lack of profit is set to decrease open data initiative funding and data maintenance as an inevitable consequence, leading to what is known as the 'benefit paradox,' whereby despite high data publication volume, open data will not generate enough profit to support its sustainability. Without adequate gains, the publishing organization will lose motivation and will hold back on expending effort on data governance beyond its internal organizational use.

Even the slightest investment in open-data quality can have a significant impact. According to international rankings such as the ODB and GODI, the UK is acknowledged as a pioneer in OGD worldwide. Furthermore, it is a leading member of the OGP. However, even with the UK's commitment to its open-data strategy, it is still susceptible to flaws. A study by Wang et al. [28] —found that the UK is practicing what is known as ''impression management,'' which gives government departments an easy and low-cost way to claim their OGD credentials by publishing inert internal management data. An analysis of the UK's open data found that the public is interested only in a few high-quality publications covering topics that matter to them. One participant described the UK's OGD program as a 'cottage industry,' which means that 'it relies on the enthusiasm, skill, and goodwill of a few resourceful government employees' to operate effectively.

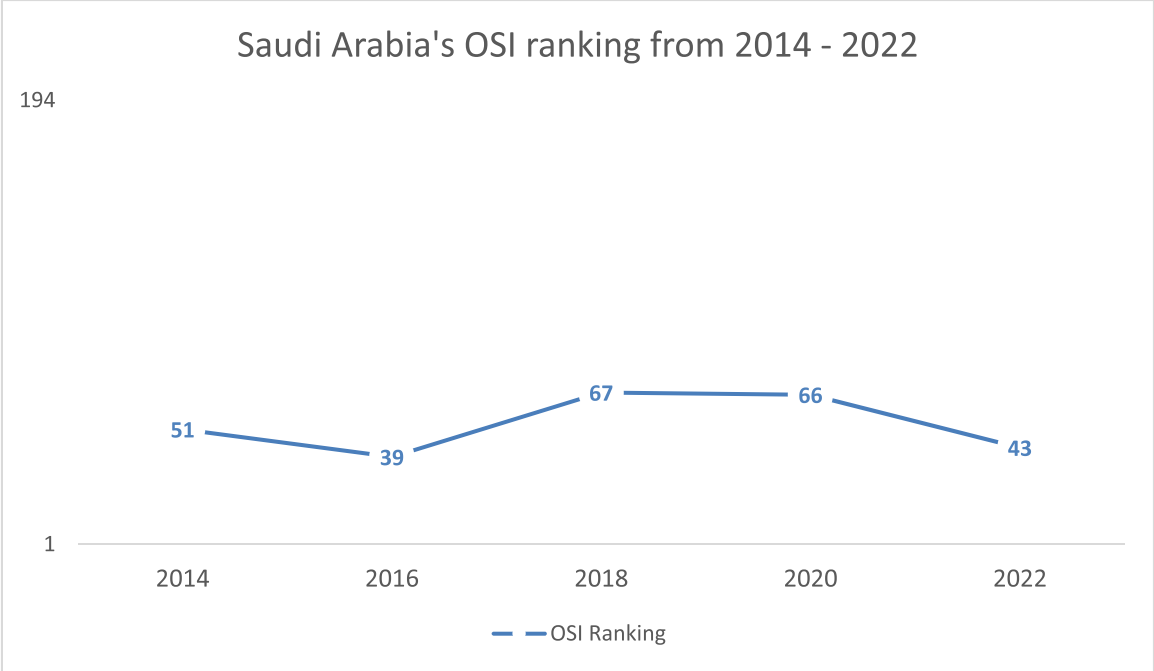

FIGURE 2. Saudi Arabia's OSI ranking from 2014 – 2022.

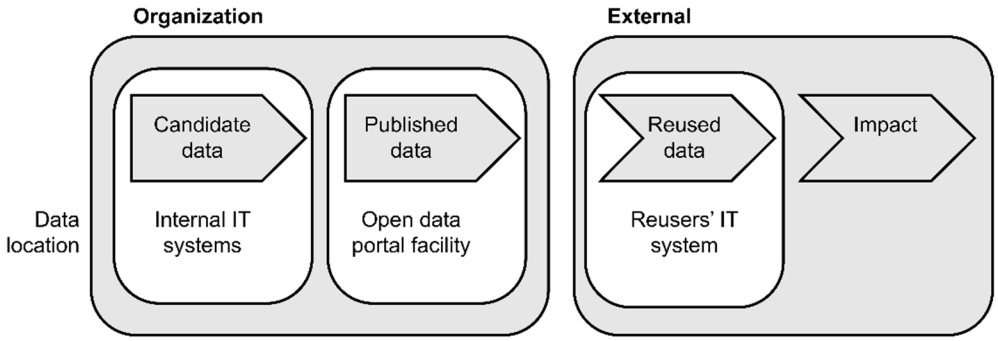

FIGURE 3. Open data impact process [27].

Overall, monitoring performance is the key to maintaining OGD sustainability. Performance management can be divided into four phases: Monitor, Analysis, Action, and Review (MAAR) [29]. After monitoring the performance of the Saudi OGD initiative and analyzing its reason, which is associated with the data quality, the next step will be to take action and review the results. Taking action to solve the issue with the open data porta's content can begin by ensuring the quality of the published data. Saudi Arabia's open data portal lists the quality criteria it complies with in its ''open data quality. standards guideline [30].'' However, because the low impact indicated undetected quality shortcomings, the solution would be to improve the quality-assessment approach.

D. FRAMEWORKS AND MODELS FOR MEASURING OPEN GOVERNMENT DATA QUALITY

Research on OGD quality examined it on a high level by evaluating the portal and its available properties or on a deep level by investigating published datasets and the quality of the data they contain. The following section analyzes the two quality assessment approaches to help elect a framework compatible with the needs of the Saudi open data portal.

1) HIGH-LEVEL FRAMEWORKS AND MODELS

The quality of OGD portals can be assessed by comparing the portal's compliance with open data standards or using an assessment model to determine the portal's fulfilment of quality requirements. Table 2 summarizes the research papers that evaluated open data portals globally. The findings of studies that included the Saudi portal are discussed in Section II-B3. The findings of the frameworks were consistent in pointing to the portal's shortcomings and eliminating the need to find a new framework at a high level.

2) DEEP-LEVEL FRAMEWORKS AND MODELS

In addition to high-level quality assessment frameworks, another approach is to evaluate the portal's published data. The deep-level quality assurance level is the focus of the

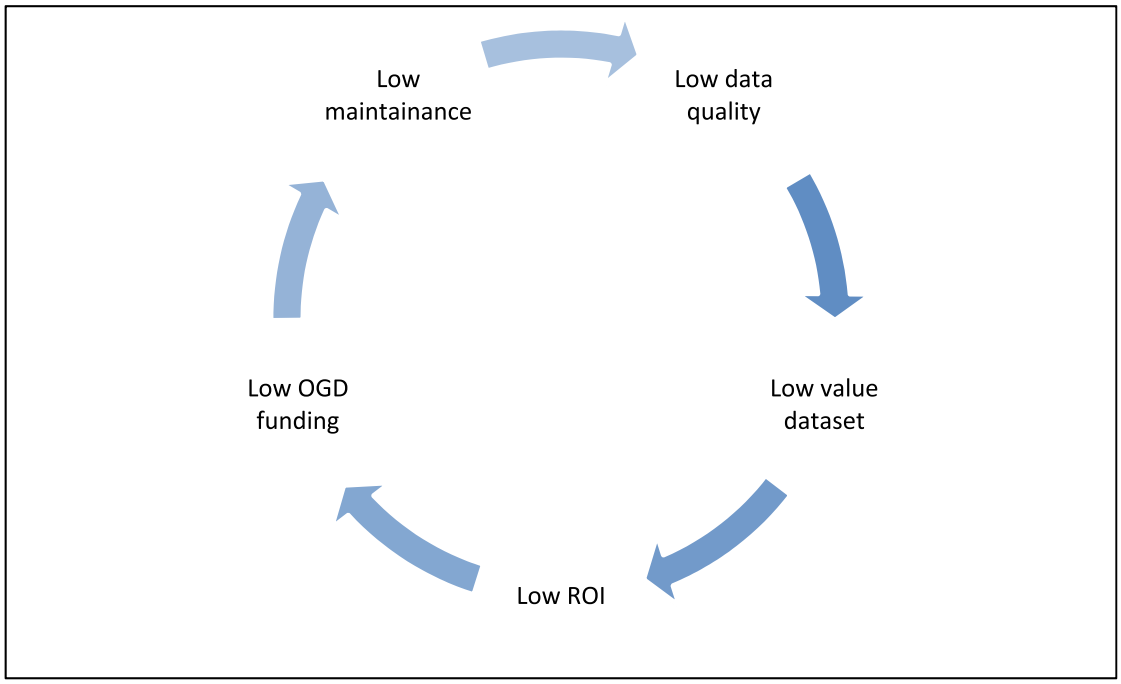

FIGURE 4. Low data quality effect cycle.

Saudi open data portal in its quality criteria [30]. This section summarizes studies that have utilized deep-level frameworks to examine their compatibility with the Saudi portal. The first two frameworks are recommended by the portal's ''Open Data Quality Standards guidelines [30]''.

1. The International Monetary Fund's (IMF) Data Quality Assessment Framework evaluates the data using five main dimensions. Each dimension comprises several elements with relevant quality indicators to track its fulfillment.

Evaluation Criteria:

- Prerequisites of quality: Providing an environment that supports statistics and resources commensurate with the needs of statistical programs.
- Integrity: Policies and practices are transparent and are guided by professional principles and ethical standards.
- Methodological soundness: data collecting methods accord with internationally accepted standards.
- Accuracy and reliability: Source data provide an adequate basis for compiling practical statistics, which are ensured through regular assessment and validation.
- Serviceability: Providing statistics with adequate periodicity and timeliness, consistent over time, and following a regular revision policy.
- Accessibility: data and metadata are easily available with adequate assistance to users.

Compatibility limitations with the Saudi open data portal: The framework primarily handles statistical data. However, the Saudi portal does not always include statistical data. For instance, the General Commission for Survey published a Portable Document Format (PDF) document named ''KSA Official Map-Arabic.pdf.'' The framework assumes formatting compliance by default and thus does not handle this quality shortfall.

2. A framework developed by Vetrò et al. [37] evaluates data quality at the dataset and cell levels. It quantifies data quality according to the following characteristics:

Evaluation Criteria:

- Traceability: having metadata associated with creating and updating the dataset.
- Currentness: the ratio between the supposed update and the delay of publication, in addition to the percentage of rows with current values
- Expiration: The ratio between the publication delay after the dataset expires and the supposed update.
- Completeness: the percentage of complete rows and cells in the dataset.
- Compliance: The percentage of the columns representing information with universal standards and the degree of the dataset's compliance with the e-Government Metadata Schema (eGMS) and Berners-Lee's five-star standards.
- Understandability: The percentage of columns with metadata was presented in a machine-readable format.
- Accuracy: percentage of cells that contain values that comply with the domain and type of their dataset. The ratio of the error of aggregation to the scale of data defines the accuracy of aggregated information.

Compatibility Limitations With the Saudi Open Data Portal: This framework is viewed as the best practice when

objectively evaluating open data, as it uses a mathematical equation to calculate a broad set of quality characteristics. However, three incompatibilities affect its applicability to the Saudi Open Data Initiative:

a) Two of the 14 criteria are dedicated to traceability. This characteristic is not required in the Saudi open data portal, and there are no enforced measures to ensure that data re-users know the list of updates and the date of the dataset's creation.

b) By requiring the percentage of errors of aggregation as a criterion, the model assumes that the published data will be fundamentally detailed, raw, and numeric, which is not the case, considering the nature of the published Saudi datasets that mainly take the form of reports, in which the accuracy of aggregation does not apply to all Saudi datasets.

c) The calculation of publication delay is based on the assumption that the date of information availability differs from the date of publication. However, in the Saudi portal, re-users do not know the date when the open data became available from the source to differentiate it from the publication date on the portal, and the date of information availability is mainly the date of publication, which renders it an unreliable measure for assessing the Saudi OGD initiative's quality.

3. Sánchez et al. [38] developed a European web-based tool, tabular data quality Assessment and Improvement of Health (TAQIH), to measure and improve the quality of healthcare datasets.

Evaluation Criteria:

- Completeness: This tool calculates the number of missing values in a sample as a whole and in a variable.
- Accuracy: Outlier detection can result from poor data collection.
- Redundancy: variables with duplication or high coloration indicate redundancy.
- Readability: a visual interface for the open data to ensure inclusive comprehension.

Compatibility Limitations With the Saudi Open Data Portal: The tool lacks key quality measures, such as timeliness. Thus, if the dataset contains outdated data, the tool will not detect it. Time is a paramount quality criterion for the Saudi portal, as its quality guideline states, making this tool inapplicable to the portal.

4. In a study by Yi [39], the authors compared the quality of UK, US, and Korean open-government site datasets.

Evaluation Criteria:

- Machine readability: checking the data format using the five-star ranking system.
- Completeness: determining whether any data are missing or inaccurate.

Compatibility Limitations With the Saudi Open Data Portal: The framework lacks key quality measures according to the Saudi open data portal, such as timeliness.

5. The Framework proposed by Nikiforova [40] applied different semantic and syntax considerations to assess data quality. Open data were examined at different levels (field, attribute, database, and dataset scope). Nine datasets were manually analyzed to determine each attribute's quality requirements and the error rate percentage in the dataset scope.

Evaluation Criteria:

- Completeness: checking for missing values in cells and missing information fields in the database.
- Correctness: ensuring that each cell has a valid value, for example, the websites have working links and the phone numbers are correct.
- Accuracy: detecting any anomalies in the data that can be viewed as outliers.
- Consistency: checking the value consistency within a single attribute in addition to the naming and standards of the fields among different databases.

Compatibility Limitations With the Saudi Open Data Portal: The framework lacks key quality measures according to the Saudi open data portal, such as timeliness.

6. Fan and Zhao [41] presented OGD quality as a quantitative equation, calculated using weighted quality variables.

Evaluation Criteria:

- Primary: checking whether the data are processed or raw.
- Update frequently: check the dataset's description for the supposed frequency of updates.
- Machine-readable: the dataset's format.
- Usage: the number of downloads.
- Timely: checking whether the dataset is updated in real-time or not.
- Completeness: the extent to which data are published.

Compatibility Limitations With the Saudi Open Data Portal
Although the framework covers the main quality issues with Saudi open data, each variable is evaluated based on the evaluator's opinion, which renders it relatively subjective. An example is the completeness variable. In the framework, the variable is given a score from to 1-5, where 5 indicates that the publisher shares all the data and 1 indicates that the publisher shares a 'very small amount of data. ' Defining the term 'very small' is subjective.

III. METHODOLOGY

The literature analysis demonstrates a problem with Saudi Arabia's present system for open-data quality assurance, which results in undiscovered quality flaws. This also highlights an issue in the absence of a suitable replacement framework. The proposed solution suggests a quantitative quality assessment framework based on the characteristics of the Saudi open data portal. Adopting a new quality assurance procedure should help identify quality defects in Saudi open data that might go undetected. The construction of a customized quality assessment framework for the Saudi Open Data Portal will undergo three stages. The first is the process of selecting the quality characteristics to identify high-quality data. Next, the factors were transformed into formulas to

TABLE 2. Research on the quality of open data portals.

Evaluation by portal comparison		
Framework's objective	**Evaluation dimensions**	**Metrics**
The authors compared the open data portals of Taiwan, the UK, Denmark, Colombia, Finland, and Saudi Arabia [26].	1- Availability: Does the data exist in any form online or offline?	Yes/no
	2- Accessibility: is the data publicly available?	Yes/no
	3- Format: is the data stored in a computer or any digital form?	Yes/no
	4- Cost: is it free of charge?	Yes/no
	5- Machine readability: can the data be easily structured by a computer?	Yes/no
	6- Update: is the data up to date?	Yes/no
	7- License: is it open-licensed?	Yes/no
	8- Social media: does the national portal use social media tools?	Yes/no
	9- Application Programming Interface (API): is API enabled in the open government data?	Yes/no
	10- Mobile application: does the government portal have a mobile app?	Yes/no
The author compared the open data portals of UAE, Bahrain, Tunis, Oman, Qatar, United States of America (USA), and the UK [31]	1- Open data policies: the documented guidelines published on the portal.	List the published policy title and its key points
	2- Open data license: permitting the use of data, with consideration of any boundaries associated with it.	List the license mane and its key points
	3- Availability: assuring that data from different groups and sectors are available.	List the portals' data groups, the available data format, and the services available in the portal
	4- Accessibility: the availability of information about the dataset that facilitates its discovery and access.	List the metatags of the datasets – list the ways that the user can search, sort, and filter the results.
	5- Encourage data using: the level of communication with the public.	List the feedback ways and the contacts available at each portal.
The paper compared Thailand's OGD website to the top six countries on the global data index, which are Taiwan, the UK, Denmark, Colombia, Finland, and Australia [32].	1- OGD websites' features: Usability, Interaction, informativeness, accessibility, and inspiration.	A table that marks a ✓ for every available feature in the countries' portal. (Yes/No)
	2- OGD websites' content: number of datasets, format, major type of data (the category with the most datasets), most viewed datasets, and languages.	A table that list the contents' details for each country.
	3- OGD websites' technologies: web server type, Operating System, CMS/DMS, programming language, and search technology.	A table that list the technical details for each country.
Comparison between the open data portals across the GCC countries (UAE, Oman, Bahrain, Saudi Arabia, and Kuwait) [21].	1- Introducing OGD: the availability of a definition, a web page with published datasets, policies, license, strategy, and guidelines	Yes/no
	2- Format: this includes evaluating timeliness, machine-readability, readability, and cost-free.	Yes/no
	3- Type of files: did the national portal provide guidelines, reports, audio, visual data, statistics, and presentations?	Yes/no

TABLE 2. *(Continued.)* **Research on the quality of open data portals.**

	4-	Services: the user should easily search the portal, sort and filter the results, find special-need services, and find the datasets divided by sectors. The users should be able to comment, rate, view print or download, see social media linking, RSS, survey, request ability, the portal should provide smartphone apps, and support different languages.	Yes/no
	5-	Connectivity: the government portal should provide a phone number, e-mail, a form-filling to help the citizens connect with them.	Yes/no
Comparison between the OGD initiative in Saudi Arabia and Brazil [20].	1-	Data availability: are there any government procurement contracts made available?	The number of published contracts.
	2-	Data openness: to what extent is detailed information available for the public?	List the attributes of the contract.
	3-	Data analytics: are the datasets analysable, and are they in a ready-to-use format?	The format of the data and the availability of visual analysis tool in the portal
	4-	The application within the government website: are there any applications that allow the citizen to report problems or make a suggestion?	The availability of social media platform.
The author wanted to compare the OGD portals amongst six middle eastern countries that are not part of the GCC, so she examined the portals of Cyprus, Turkey, Egypt, Iran, Lebanon, and Jordan [33].	1-	The total number of datasets.	List the number of datasets in each portal.
	2-	The availability of links to external sites.	Yes/No
	3-	The number of data categories.	
	4-	Timeliness: based on the last update on the dataset is the data up to date or not?	Yes/No
	5-	Metadata availability: the availability of basic information to describe the data.	Yes/No
	6-	Data format.	List the data format available in each portal (CSV, XML, PDF…etc.)
	7-	Searchability: the availability of a search box.	Yes/No
	8-	Social media plugins: the portal incorporation of social media to help engage with users.	Yes/No
	9-	Social media availability: the users' ability to activate social media portals.	Yes/No
	10-	Data visualization: the users' ability to conduct and onsite analysis.	Yes/No
An evaluation of Taiwan's open data platform based on a scoring system [34].	1-	Security: Is there an information security management mechanism?	Score based on the standards of the administrative Yuan (0: there are no security policy – 5: 50% of information security is operated – 10: 75% of information security is operated – 15: the security is fully operated – 20: passed a certified information security standard like ISO)
	2 – Accessibility		
	2.1 Limits on using the data		Score (0: not open to use – 5: use with limitations – 10: open without limitations)

TABLE 2. *(Continued.)* **Research on the quality of open data portals.**

	2.2 Readability using API.	Score (0: automatic reading language or API format is unavailable – 5: the percentage of datasets with API is between 0% and 50% - 10: the percentage of datasets with API is between 50% and 75% - 15: the percentage of datasets with API is between 75% and 100%)
	2.3 ODB: According to the 2015 ODB, there are 15 categories for the open datasets, how many datasets in the portal match their categories.	Score (0: none of the datasets match – 5: between 1 and 5 categories are found – 10: between 6 and 10 categories are found – 15: between 10 and 14 categories are found – 20: all 15 categories are found)
	2.3 Searchability	Score (0: there is a search button but it can't find any matching results – 5: irrelevant datasets appeared in the results – 10: between 6 and 10 categories are found – 15: retrieve relative datasets by typing key words– 20: the portal uses word association for better search results)
	2.4 Format: the score is based on the Berners-Lee 5-star principle.	Score (0: there are no downloadable datasets in the portal – 5: datasets are available for reading in PDF or JPEG – 10: structured data like EXCEL files – 15: database format like CSV or XML– 20: machines can access, save and apply all data in the dataset– 25: the portal has linked data)
	2.5 License-free	Score (0: no regulation about open license – 5: there are regulation on a certain extent – 10: the data is open licensed)
	3 – Quality	
	3.1 Primary: the availability of raw data.	Score (0: raw data is unavailable – 5: raw data is available)
	3.2 Timely: the availability of information about the time of data collection, last update, and frequency of update.	Score (0: none of the 3 descriptions are included – 5: one of three descriptions is available - 10: two of three descriptions is available - 15: all three descriptions are available)
	3.3 Accuracy: metadata description matches the content.	Score (0: the description is false– 5: the description is correct)

TABLE 2. *(Continued.)* **Research on the quality of open data portals.**

	3.4 Integrity: based on the standard regulation, there are 22 categories issued by the National Development Council in 2015.	Score (0: between 0 and 5 categories are a match – 5: between 6 and 10 categories are a match – 10: between 11 and 15 categories are a match – 15: between 16 and 20 categories are a match – 20: over 20 categories are a match)
	3.5 Abundance: use analysis can be conducted easily	Score (0: no assistance is provided for the user – 5: data can be presented in rows and columns – 10:data are represented visually)
	4 – Other functions	
	4.1 Discussion: The ability for the user to give feedback.	Score (0: users can not give feedback – 5: provide remark or feedback mechanism- 10: provide forums- 15: provide data communities)
	4.2 Scoring and ranking	Score (0: scoring or ranking are not available–5: scoring or ranking are available)
	4.3 The demand for unpublished data	Score (0: there is not a data-demanding application – 5: there is a data-demanding application – 10: there is a data-demanding application that responds)
	4.4 The users can view the number of downloads and visits for categories	Score (0: the user cannot see the number of downloads nor the number of visits – 5: the user can see either the number of downloads or visits – 10: the user can see the number of downloads and visits)
	4.5 number of downloads and visits for datasets	Score (0: there is not a total number of downloads available – 5: user can see either the number of downloads or visits – 10: the user can see the number of downloads and visits)
The author evaluated India's OGD portal using Total Quality Management (TQM) model [35].	1- Enablers: there are data contributors that supply data – leaders in the government – personal to work on the open data – infrastructure – processes to coordinate 2- Drivers: re-users/stockholders - suppliers – community – peer pressure among public agencies – economic condition – market condition – the environmental condition 3- Results: datasets quality – user satisfaction – performance benchmarking – networking among agencies.	The evaluator should be able to identify what represents each indicator in their local OGD portal.
Evaluation of "the openness" of the UK's open data portal using	1- Granularity: the number of datasets that are not aggregated (not a report or summary or category)	The number of aggregated datasets found in the sample and their percentage.

TABLE 2. *(Continued.)* **Research on the quality of open data portals.**

what is the authors defined as "the ordinary citizen test" performed on a sample of the datasets [36].	2- Timeliness: the number of datasets that are less than 30 months old.	The number of outdated datasets and their percentage.
	3- Machine-readable: the number of datasets that scored 2-Stars or higher in the Berners Lee scale.	The number of unstructured datasets and their percentage.
	4- Non-proprietary: the number of datasets that are provided in an open software format.	The number of datasets that are not in an open format and their percentage.
	5- License-free: the number of datasets that did not require permission or filling forms to access them.	The number of datasets with license restrictions and their percentage.
	6- Operable/processable: the number of datasets that did not fail to function.	The number of datasets that fail to function or were inoperable and their percentage.
	7- Published: the number of datasets that are available compared to the datasets that are announced but were not found.	The number of datasets without entries and their percentage.
	8- Cost-free: the number of datasets that did not ask for fees.	The number of datasets that required payment to access it and their percentage.
A benchmarking usability evaluation framework that focuses on the users experience while navigating the OGD portal [25].	1 – Open dataset specification: 1.A description of dataset: how the data was collected and for what purpose. 1.b Publisher of datasets: the organization that published the dataset. 1.c Thematic categories and tags: addressing the main topics that the dataset covers by identing the categories keywords associated with it. 1.d Release date and up to date: identifying a specific time period for the dataset i.e., data update frequency, date of publication. 1.e Machine-readable formats: presenting the dataset in a format that facilitates its reusability. 1.f Open data license: providing licence information related to use the published datasets. 1.g Visualization and analytics tools: the ability to present the data in charts or maps.	Three-point scale to measure the successful completion of a task where 1=unfulfilled, 2 = partially fulfilled, and 3= done.
	2 – Open dataset feedback 2.a Documentation and tutorials: to help the users in learning how to use the portal. 2.b Forum and contact form: the ability to submit feedback and having a forum for the users to discuss publicly. 2.c User rating and comments: enabling user participation. 2.d Social media and sharing: integration with social media to allow feedback sharing *3 – Open dataset request* 3.a request form: allowing the request and suggestion of new open data. 3.b List of requests: for the users to see their requests and their state 3.c Involvement in the process: allowing the involvement in the active request i.e., show interest in the same datasets	

TABLE 3. A list of the most common OGD qualities.

Dimension	Definition	Level
Completeness	Representing all the relevant data with no missing values.	Cell-level
Timeliness	Data are updated frequently.	Dataset level
Primary/Granularity	Data are not aggregated or modified.	Dataset level
Accuracy/Correctness	The likelihood that the data reflect reality.	Cell-level
Machine-readability	Data are provided in a format that the machine can process.	Dataset level
Consistency/compliance	The data comply with the standards within the dataset, and the data value does not change between different versions of the dataset.	Cell-level
Readability /Usability	The data are clear, and ordinary users with no technical background can understand it easily.	Cell-level
Usefulness/Value	The data are valuable, and the users can gain an advantage from it.	Dataset level
Redundancy/Uniqueness	Avoid duplicating the data or representing the information in unnecessary detail.	Cell-level

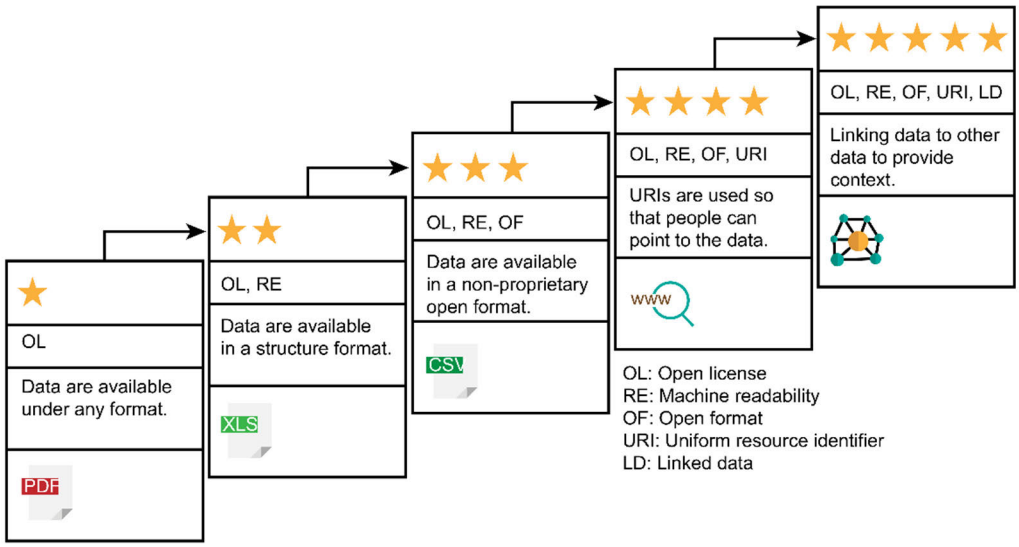

FIGURE 5. Berners-Lee's five-star open data-scoring system.

produce a score that reflects the dataset's quality. The last phase proposes weighted scores for the framework to reflect the importance of specific quality criteria based on Saudi experts' views.

A. FIRST STAGE: SELECTING THE QUALITY CHARACTERISTICS

An examination of the open data literature demonstrates how to disparate the criteria for high-quality open data. Table 3 presents the most frequently used deep-level quality metrics. The metrics were selected based on a literature review of the study. The criterion needs to fulfill the following conditions:1) it should assess the quality of the open data based on the dataset or its values, not the portal itself; 2) the metric must be mentioned more than once; and 3) it must be objectively measurable. 3) The metric must be applied to the Saudi open data portal; for example, traceability is disqualified because the portal does not enforce it. The findings of Šlibar et al.'s [42] study confirm that the selected criteria are most frequently used in the literature on open data. Quality

TABLE 4. Quality weight scores.

Weight	Score	Abbreviation	Criteria
0.83	83	CLC	Cell-level completeness
0.84	84	ILC	Information-level completeness
0.91	91	G	Granularity
0.89	89	PT	Publication Timeliness
0.88	88	CT	Content Timeliness
0.91	91	MR	Machine Readability
0.83	83	CON	Consistency
0.81	81	ACC	Accuracy
0.96	96	MD	Metadata
0.94	94	CG	Comprehensive Format
0.69	69	U	Usage
0.66	*66*	*CD*	*Column Duplication*
0.69	*69*	*RD*	*Row Duplication*
0.675	67.5	D	Duplication

TABLE 5. Quality scores of the proposed framework.

	CLC	ILC	GR	PT	CT	MR	CON	ACC	MD	CF	U	D	TOTAL
	0.999	0.218	0	0	0	0.6	1	0.97	0.54	1	0.7	0.91	**57.8%**

TABLE 6. Quality score after weighting the result.

CLC	ILC	GR	PT	CT	MR	CON	ACC	MD	CF	U	D	Total
0.82	0.18	0	0	0	0.54	0.83	0.78	0.1	0.94	0.483	0.61	**47.747%**

measures that do not violate selection standards are integrated into the suggested framework.

B. SECOND STAGE: FORMULATING QUALITY MEASURES

The proposed framework in Appendix A quantifies the elected nine quality characteristics as follows:

1) Completeness: The completeness of the dataset is measured on two levels: The first is the cell level; this indicator measures the completeness based on the number of empty cells or values like (NULL, '-') that effects the dataset quality. The second level is information. Unnecessary reticence can be measured by comparing the published dataset with the original dataset in the organization's systems to consider unpublished non-private data. Private data (e.g., names, Social Security numbers, addresses, etc.) are the only data that should be kept from publishing and are fixed to present a certain narrative; they cannot be analyzed or input in useful computations by data scientists and researchers. For this reason, the more granular or primary the data, the higher the dataset's quality.
2) Timeliness: A dataset was inspected to determine whether it contained recent or outdated data. The currentness of the dataset was examined at two levels. The first level is the dataset's publication timeliness in the context of its updates,[1] and the second level is the freshness of the data within the dataset. Portal visitors can notice that publishers upload old datasets

[1]this criterion is designed to fit the nature of the Saudi data publication frequency, if the dataset is updated daily then the algorithm will give a division by zero error, and for frequently updated data the expiration value will be greater than the update frequency value with the absence of delay we will have a quality value that is larger than 1, but these are special cases that do not occur in the Saudi portal since the most frequently updates datasets are updated quarterly.

TABLE 7. Quality scores of the framework by Vetrò et al. [37].

Track of creation	Track of update	Percentage of current rows	Delay in publication	Delay after expiration	Percentage of complete cells	Percentage of complete rows	Percentage of standardized columns	eGMS compliance	5star open data	Percentage of columns with metadata	Percentage of columns in a comprehensible format	Percentage of accurate cells	Accuracy in aggregation	Total
tc	tu	pcr	dp	dae	pcc	pcpr	psc	eGMS	5star	pcm	pcuf	pac	ea	
0.333333	0.5	0	0.135	0.8361	0.99981	0.99	0	0.64	0.6	0	1	1	1	57.3 %

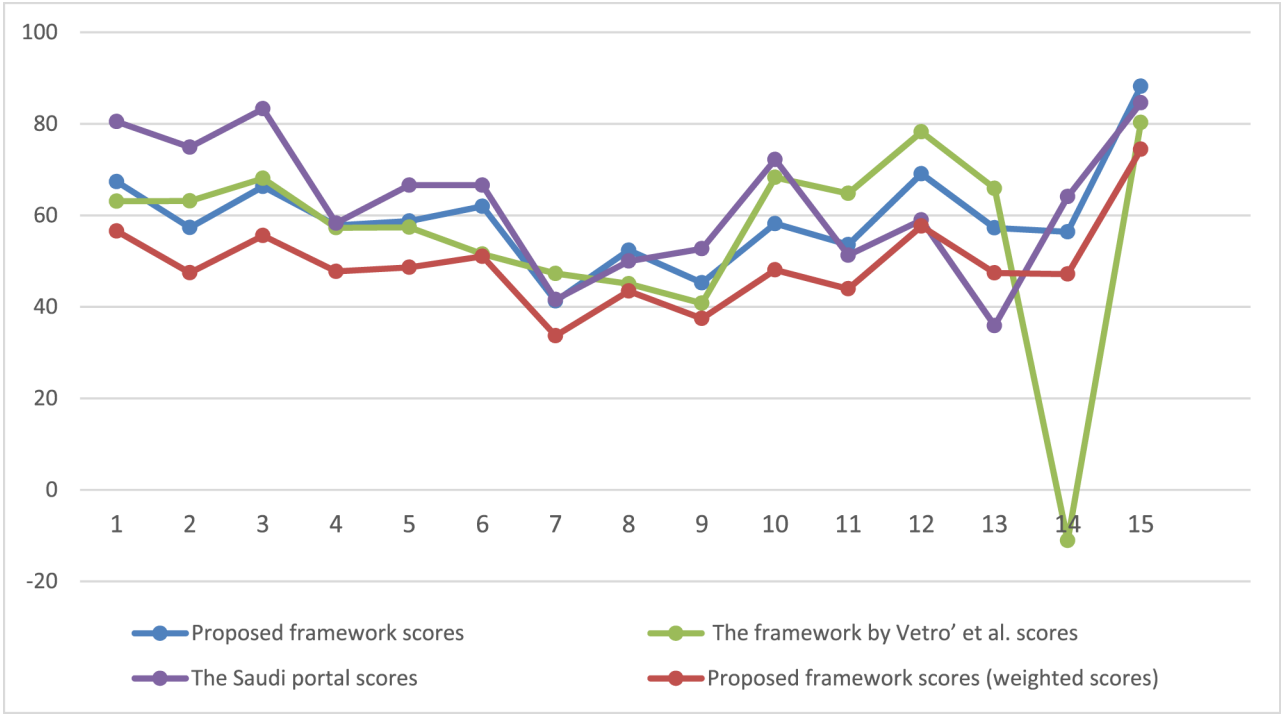

FIGURE 6. Comparison between the 4 quality assessment results.

and update datasets frequently without changing their content. By the time of writing this document, only 13.23% of the publishers committed to providing data as recently as the period from 2019-2020.

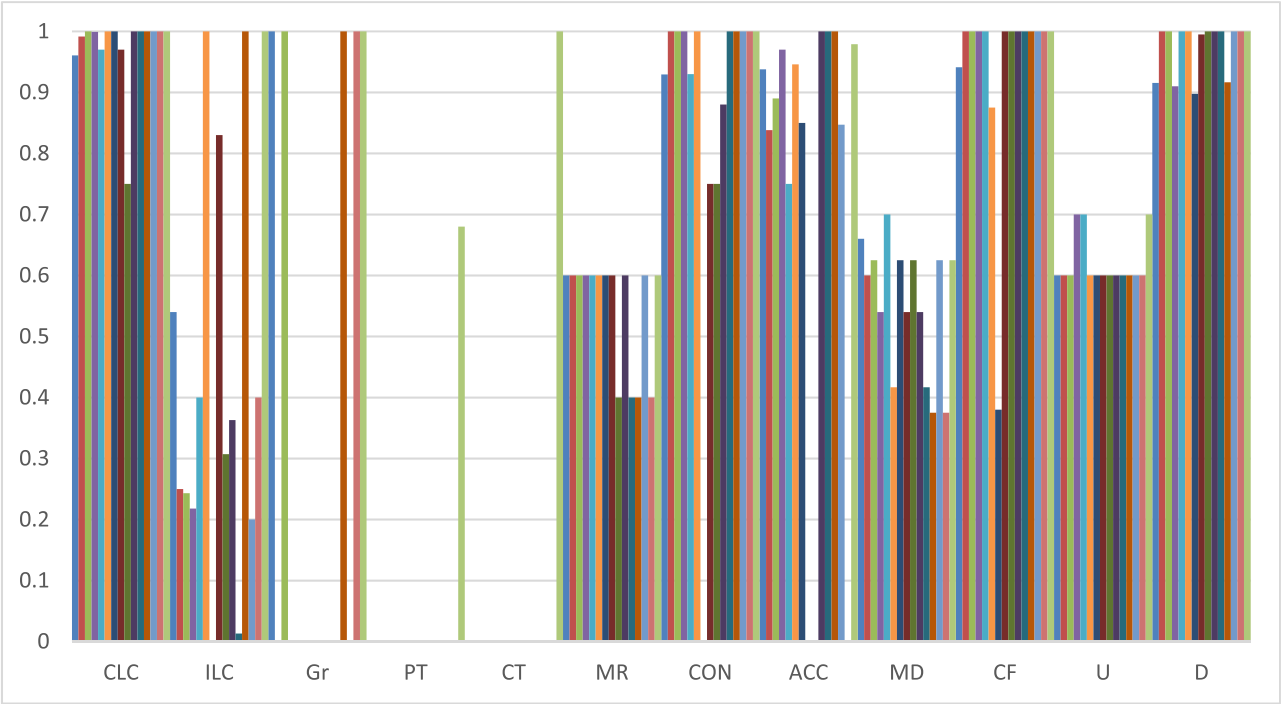

FIGURE 7. The assessment results of the proposed framework.

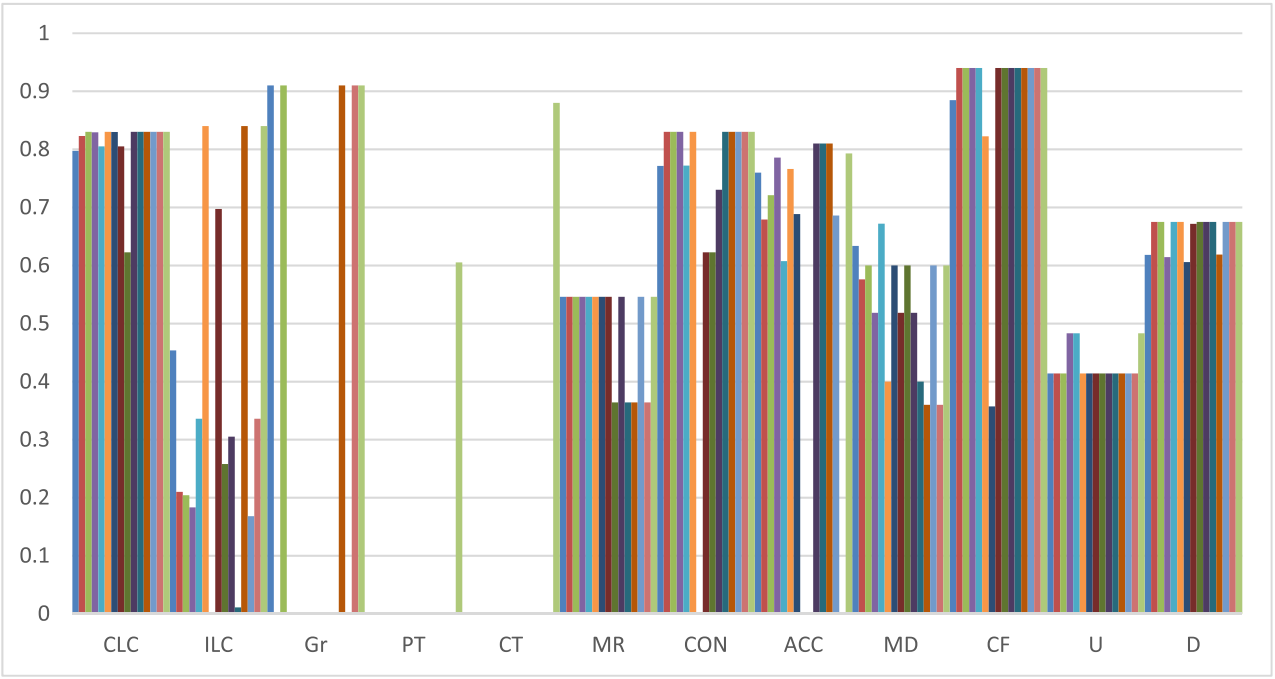

FIGURE 8. The assessment results of the proposed framework (after weighing the score).

3) Machine reusability: measured by scoring the dataset based on its available format using Berners-Lee's five-star open data scoring system [43], as shown in Fig. 5.
4) Consistency: The dataset is consistent when the values in the columns with standards follow the same patterns, e.g., the data format for dates can be 'yyyy/mm/dd' or 'yy/mm/dd' but not both. Another example is the text in a numerical column. Note that some columns, e.g., 'notes,' cannot be standardized.
5) Accuracy: Extreme values in datasets that do not fit within the normal ranges can indicate poor data entry or collection methods. To detect these values, it is

FIGURE 9. The assessment results of the framework by Vetrò' et al. [37].

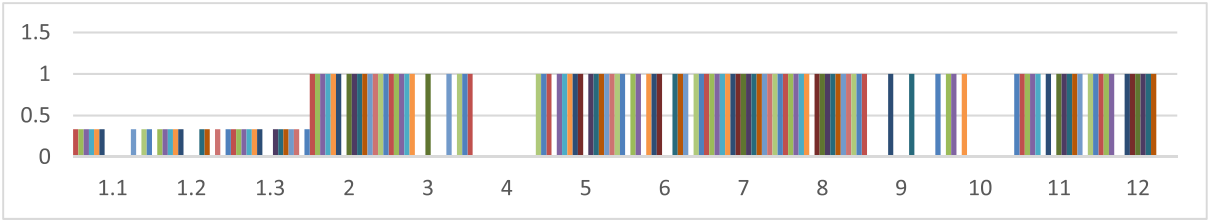

FIGURE 10. The assessment results of the Saudi portal's model.

necessary to have a univariate outlier detection method that individually tests each column and provides a more accurate test than measuring the values within the dataset as a whole. Given the common structure of the Saudi open dataset, the Tukey test was considered a compatible measure. The Tukey test calculates the mean of each column (Q2), the mean of the values above the mean (Q3), and the mean of the values below the mean (Q1). A formula was then applied to obtain the range for which any values that are outside this range were considered outliers. Appendix B.1 demonstrates the detection of outliers using RapidMiner Studio.

$$[Q1 - 1.5\,(Q3 - Q1)\,,\, Q3 + 1.5(Q3 - Q1)] \quad (1)$$

6) Understandability: Users' ability to interpret and understand the meaning of data is affected by metadata availability. Appendix C contains a list of the required metadata based on the recommendations of the Working Group Metadata Cooperation OGD Austria [44], listed in the form of an examination sheet. The second factor affecting the dataset's understandability is the comprehensive format, as data gathered from the dataset are often stored automatically or converted from one format to another, which can cause data to be published in a format that does not make sense to the reader. Unfortunately, when these mistakes are unnoticed by a publisher, users of the data end up with nonreusable datasets.
7) Use: The Saudi open data portal uses a five-star rating system that allows the public to rate the datasets and uses the average of the ratings as a score for the dataset. The rating system's default value of the rating system was set to zero stars, which indicates the public's use of the published dataset.
8) Redundancy: If a row or column is repeated more than once, this can indicate poor database management that affects the quality of the dataset. Appendix B.2 shows how to utilize RapidMiner to measure this factor.

After calculating all metrics, 12 quality scores were obtained. The total quality of the dataset shall then calculated

TABLE 8. Quality scores of the portal's framework.

Requirement	Checklist score
1) Machine readability: ensures that	
1.1 Raw data should be provided in Excel or .CSV format.	0.33
1.2 Merging cells are avoided.	0.33
1.3 The data should be free from pictures or graphs.	0.33
2) There is no data leakage, meaning that the dataset does not contain private or unauthorised data.	1
3) Overall data quality	1
4) Metadata should be provided.	0
5) The data should have a primary key.	1
6) The data should be within a specific time range.	0
7) The size of the data should be less than 20 MB.	1
8) The document should be named in appropriate English.	1
9) Data should be up-to-date and timely.	0
10) The data should be valuable to help motivate entrepreneurs to provide digital products or services.	0
11) English and Arabic content should be separated.	1
12) Frequently updated and new data should be added to the same file.	0
Total	**58.33%**

as:

$$Quality = \frac{\sum QualityScore}{12} * 100. \quad (2)$$

C. THIRD STAGE: WEIGHING THE SCORE

To impose fairness on the effect that the criteria will have on the final score, Ten elected Saudi pioneers in the data analysis and data quality field were asked to rate the criteria of the proposed framework based on their importance. Selecting precisely ten participants was due to the difficulty of finding Saudi specialists with niche expertise in data science, and the sufficiency allowed for a straightforward scoring distribution. Appendix D details the participants' credentials and responses, which are summarized in Table 4. The participants rated each criterion from to 1-10, with 1 indicating that the character is not essential to the quality of the open government dataset and 10 indicating that the character is crucial to the dataset's quality. The total score assigned to each criterion was then divided by 100 to generate the weight value. The participants were asked to rate the column duplication and row duplication individually, since early versions of the framework had separate scores for each type of duplication; they were now merged under one criterion with the average score as its weight.

IV. USE CASE

This section presents a case study of local traffic accidents. In the Saudi Open Data Portal, the *General Directorate of Traffic,* affiliated with *the Ministry of Interior*, is the government authority responsible for providing traffic and accident reports. Among all the datasets published by the directorate, the most comprehensive was the "*Traffic Accident Statistics as of 1439 H.*" [45] The dataset provided a monthly report for 17 regions across Saudi Arabia that enumerated the accordance of the different attributes in detail. The dataset contained 17 sheets, each of which was processed separately using the framework proposed in Appendix A. After calculating the scores on each sheet, the average score was used as the final score. The dataset score 57.8%, as shown in Table 5.

The following quality characteristics positively affected the dataset:

1- Cell-level completeness (CLC): Only three out of 16,146 cells were empty.

TABLE 9. The final assessment results of the Saudi OGD datasets.

The Saudi portal score	Italian framework scores	Proposed framework scores (weighted score)	Proposed framework scores	N0.	Name
80.5	63.1	56.568	67.37	**1**	Salaries Scales for Civil Servants
74.9	63.14	47.4	57.33	**2**	Number of Beneficiaries of Social Security Agency From 1435 to 1441
83.3	68.09	55.58	66.3	**3**	General Education- Number of Teachers by region 1437-1440
58.3	57.3	47.747	57.8	**4**	Traffic Accident Statistics as of 1439 H
66.6	57.4	48.63	58.75	**5**	Social Development Bank Loans for 2019
66.6	51.53	51.03	61.98	**6**	food truck parking location
41.6	47.28	33.678	41.27	**7**	General Education - Number of new students by region 1437-1440
50	45.04	43.45	52.375	**8**	Essential Medicines List
52.7	40.78	37.46	45.26	**9**	List of licensed pharmacies to sell psychological drugs
72.2	68.3	48.072	58.19	**10**	Ministry of Health Budget from 1427 to 1440H
51.28	64.8	43.94	53.58	**11**	Number of Declarations in the General Authority of Zakat and Tax
58.97	78.25	57.639	69.09	**12**	Government Budget Allocation of the first chapter 2012
35.89	65.9	47.40	57.26	**13**	Saudi pilgrims companies
64.1	-11.08	47.15	56.4	**14**	Total of rainfall observed by PME MET stations in 2004
84.6	80.3	74.43	88.2	**15**	residential detaiels of land deals in all reagions November 2020

2- Machine readability (MR): The dataset scored three out of five stars in the Burner-Lee scoring system since it was provided in. XLS and.CSV format.
3- Consistency (CON): All rows of every column had the same data type and format.
4- Accuracy (ACC): Using the Tukey test algorithm in RapidMiner, only 327 anomalous cells were detected and were found to be either below or above the normal range of their columns.
5- Comprehensive format (CF): the dataset had no unreadable content.
6- Usage (U): the dataset had a 3-star rating on the website.
7- Duplication (D): The datasets had no duplicated rows but 201 duplicated columns of 1242 columns.

The following factors negatively affected the dataset's quality:

1- Granularity (G), publication timeliness (PT), and content timeliness (CT) did not exist.
2- There was a shortage of metadata (MD) because 13 of the 24 metadata points were identified based on the metadata requirements in Appendix C.
3- Information level completeness (ILC) affected the quality, with only 14 attributes published for the 64 non-private attributes in the original database. Some of the 64 unpublished attributes were the status of the vehicle, the status of the road, the name of the insurance company, and the neighbourhood where the accident occurred. These attributes can help detect patterns.

TABLE 10. Detected quality deficiencies in Saudi OGD by different quality assessment frameworks.

The quality assessment framework / the detected quality deficiency	Metadata	Traceability	Timeliness	Consistency	Accuracy of aggregation	Valuable data	Data quality	ILC	Usage	Granularity
The proposed framework	✓	*N/A*	✓	⚠	*N/A*	*N/A*	*N/A*	✓	⚠	✓
The weighted proposed framework	✓	*N/A*	✓	⚠	*N/A*	*N/A*	*N/A*	✓	✓	✓
the framework by Vetrò' *et al.* [37]	✓	✓	✓	✓	✓	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
the Saudi portal's model	✓	*N/A*	✓	*N/A*	*N/A*	✓	✓	*N/A*	*N/A*	✓
The quality criterion's compatibility with the Saudi data portal	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓

*⚠ : The symbol indicates that the framework measures the metrics without detecting a quality issue.

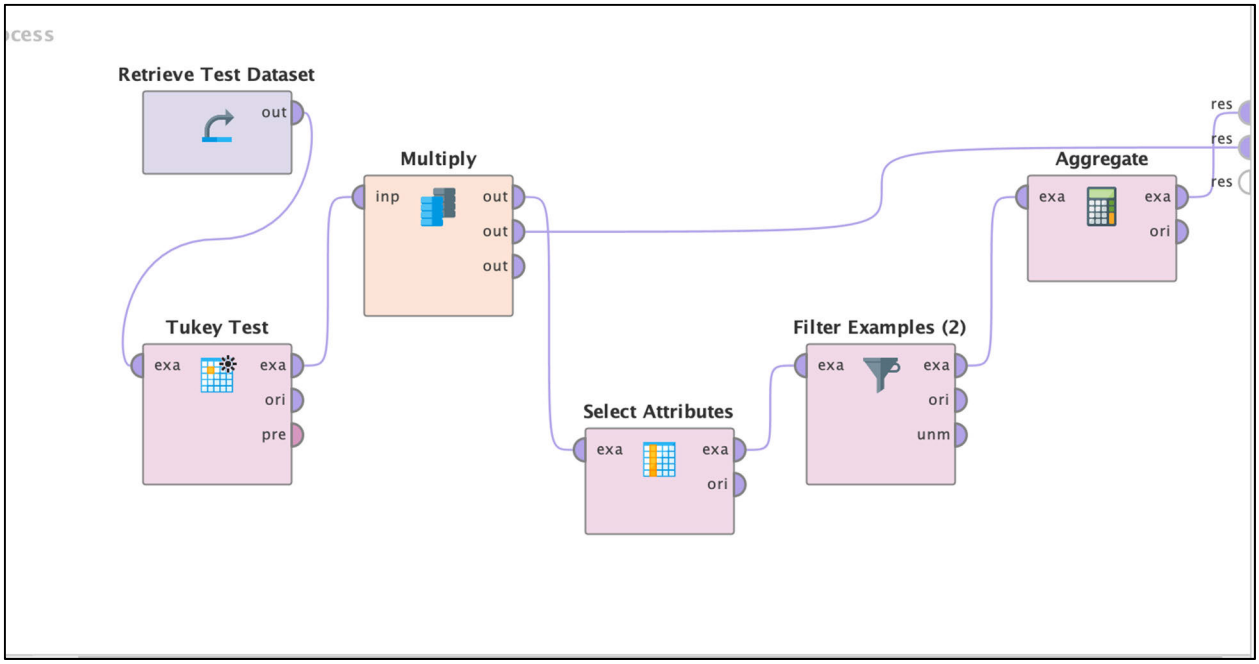

FIGURE 11. Accuracy process in RapidMiner.

The published dataset is compared to the original dataset by obtaining a copy of the original accident report that the local General Department of Traffic utilized to collect accident information. This document shows the level of detail in the original databases. When researchers can obtain such a detailed amount of data, knowledge that benefits government authorities and the public can be generated. For example, a group of researchers from King Saud University in Riyadh cooperated with the traffic department and gained access to 83,605 records and 60 attributes [46], which were displayed in the form of three tables containing information about the accident, vehicle, and accident parties. They applied data

TABLE 11. OGD quality metrics.

Quality characteristic	Description	Equation	variables	scale
Completeness	Cell-level completeness indicates the percentage of filled cells	$CLC = 1 - \frac{\text{nec}}{\text{nr} * \text{nc}}$ (Eq. 3)	nec: number of empty cells nr: number of rows nc: number of columns	[0 - 1]
	Information level completeness: the percentage of openness is calculated using the ratio of published information to the non-private information in the original dataset.	$ILC = \frac{\text{pa}}{\text{att} - \text{pra}}$ (Eq. 4)	pa: The number of published attributes from the dataset att: The number of attributes/columns in the dataset pra: The number of private attributes in the dataset	[0 - 1]
Granularity	A binary indicator to determine if the data is processed into a report or a summery rather than providing row data that can be processed	Gr = 0 if processed information are available Gr = 1 if raw data are available		[0,1]

Timeliness	Publication Timeliness: the ratio between the number of days until the dataset is expired to the delay in update.	**Algorithm 1: Calculating Datasets publication Timeliness**
		Input:
		latest provided update : dd1/mm1/yyyy1
		the day of examination: dd2/mm2/yyyy2
		next expected update : dd3/mm3/yyyy3
		UF: the update frequency of the dataset (in days)
		Output: T
		if $mmx \leq 2$ **then**
		$mmx = mmx + 12$; $yyyyx = yyyyx - 1$;
		$Expiration = (((146097*yyyy3)/400 + (153*mm3 + 8)/5 + dd3) - ((146097*yyyy2)/400 + (153*mm2 + 8)/5 + dd2))$;
		$Delay = (((146097*yyyy2)/400 + (153*mm2 + 8)/5 + dd2) - ((146097*yyyy1)/400 + (153*mm1 + 8)/5 + dd1))) - UF$;
		if $Expiration < 0$ **then**
		$EXP = 0$;
		else
		$EXP = \lfloor Expiration \rfloor$;
		if $Delay < 0$ **then**
		$D = 0$;
		else
		$D = \lfloor Delay \rfloor$;
		$T = \frac{EXP}{(D+1) * (UF-1)}$;
		(Eq. 5)
			Expiration: the time between the current date and the next update.
			Delay: Number of days passed without the expected update.
			EXP: a real rounded expiration value.
			D: a real rounded delay value.
				[0 – 1]

TABLE 11. *(Continued.)* **OGD quality metrics.**

	Content Timeliness: calculate the freshness of the data that the dataset holds within.	**Algorithm 2: Datasets content Timeliness**
		Input: the date of examination: YYYY1/MM1/DD1
		latest date provided in the data set: YYYY2/MM2/DD2
		UF: update frequency (in days)
		Output: CT
		if $mmx \leq 2$ **then**
		$mmx = mmx + 12$; $yyyyx = yyyyx - 1$;
		$NOD = (((146097*YYYY1)/400 + (153*MM1 + 8)/5 + dd1) - ((146097*YYYY2)/400 + (153*MM2+ 8)/5 + dd2))-UF;$
		if $NOD < 0$ **then**
		$NOD = 0;$
		$CT = 1 - (NOD/NOD);$
		(Eq. 6)
			NOD: the number of days between the the date of examination and the last date provided in the dataset.	(0,1)

Machine-reusability	Determined by Berners-Lee 5-star open data.	$MR = \frac{\text{fsods}}{5}$ *(Eq. 7)*	Fsods: five-star open data score	[0 - 1]
Consistency	The percentage of cells that comply to column's format standards	$CON = 1 - \frac{\text{nfc}}{\text{nr} * \text{nc}}$ *(Eq. 8)*	nfc: number of fault cells nr: number of rows nc: number of columns	[0 - 1]
Accuracy	Detect anomalies in the data using interquartile range to find the percentage of outliers.	$ACC = 1 - \frac{\text{outlier}}{\text{nr} * \text{nc}}$ *(Eq. 9)*	outliers: the values detected by HBOS nr: number of rows nc: number of columns	[0 - 1]
Understandability	Metadata: the number metadata found in the dataset based on Appendix C in perspective of their weight.	$MD = \frac{\sum_i wi * ai}{24}, \text{i} = 1,2 \ldots 16$ *(Eq. 10)*	w: weight a: attribute availability	[0 - 1]
	Comprehensive Format: The percentage of columns written in a clear readable format.	$CF = \frac{\text{ncuf}}{\text{nc}}$ *(Eq. 11)*	ncuf: the number of readable columns. nc: the number of columns in the dataset	[0 - 1]
Usage	the public's ratings.	$U = \frac{\text{fsr}}{5}$ *(Eq. 12)*	fsr: five-star rating.	[0 - 1]
Redundancy	Duplication: The percentage of duplicated rows and coulmns.	$D = 1 - ((0.5 * \frac{\text{ndr}}{\text{nr}}) + (0.5 * \frac{\text{ndc}}{\text{nc}}))$ *(Eq. 13)*	ndr: number of duplicated rows nr: number of rows in the dataset. ndc: Number of duplicated columns	[0 – 1]
			nc: number of columns in the dataset	

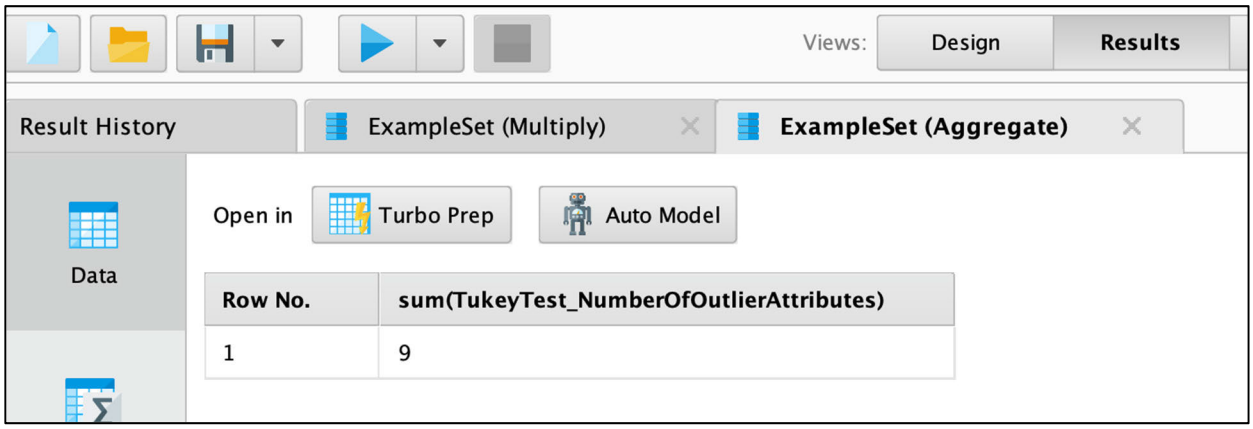

FIGURE 12. The results for the accuracy process.

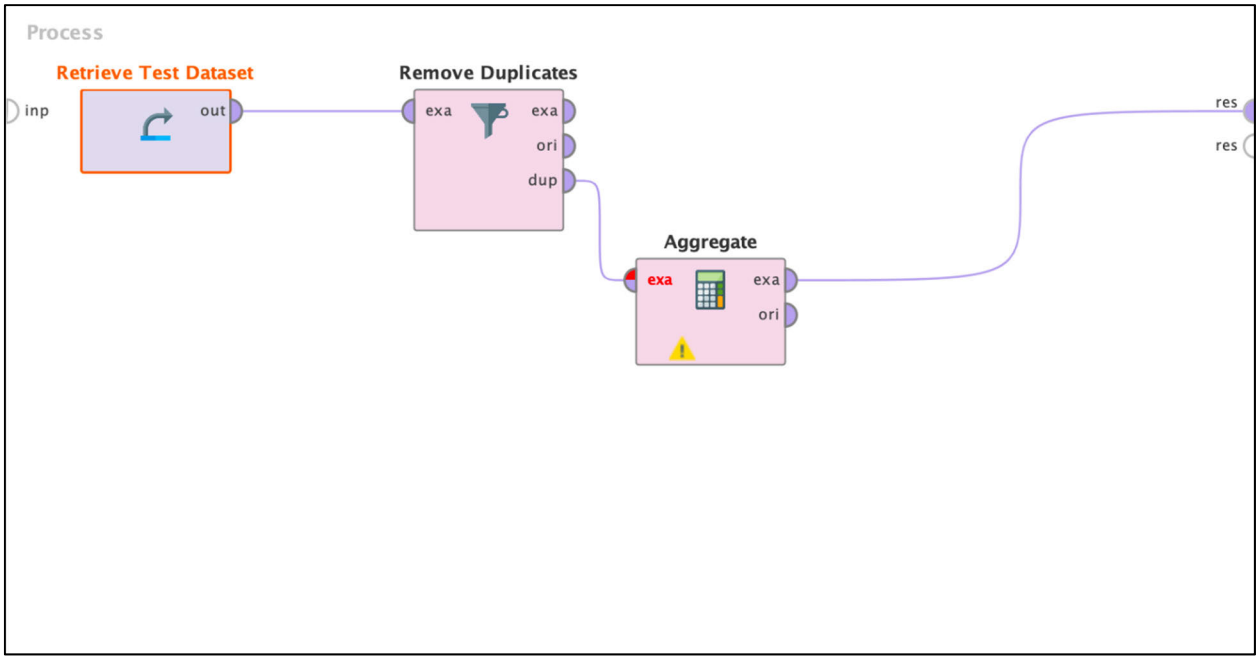

FIGURE 13. Row duplication process in RapidMiner.

mining techniques to determine that older cars are more prone to accidents than modern vehicles, and that distracted driving is the leading cause of traffic fatalities. The processing level to reach these conclusions could not be obtained with the dataset published by the traffic department in the Saudi open data portal.

Table 6 shows the evaluation results after weighing the scores using the values in Table 4. Table 6 shows how certain quality criteria like Duplication (D), Use (U), and Accuracy (ACC) are the most affected in the weighing process since the experts gave them lower importance scores compared to criteria like Metadata (MD) and Comprehensive format (CF) that kept their weight.

A comprehensive quality assessment requires measuring the same dataset using the international OGD standards. Based on the candidate's framework in Section 2.2.4, the framework by Vetrò et al. [37] was chosen for comparison because it has the most quality metrics and it inspired the structure of the proposed framework. Both frameworks use quantitative quality measures and assess the dataset at a deep level, thereby making their readings comparable. The results of the quality assessment are presented in Table 7.

The quality characteristics that positively affected the dataset were that it was examined before its expiration date and that it had only three missing cells and two incomplete rows. The dataset followed some eGMS standards by listing the creation date, source, title, and publisher. The dataset received a three-star rating in the portal and presented all the data in an understandable format. The cells had values that complied with the domain of the dataset and the type

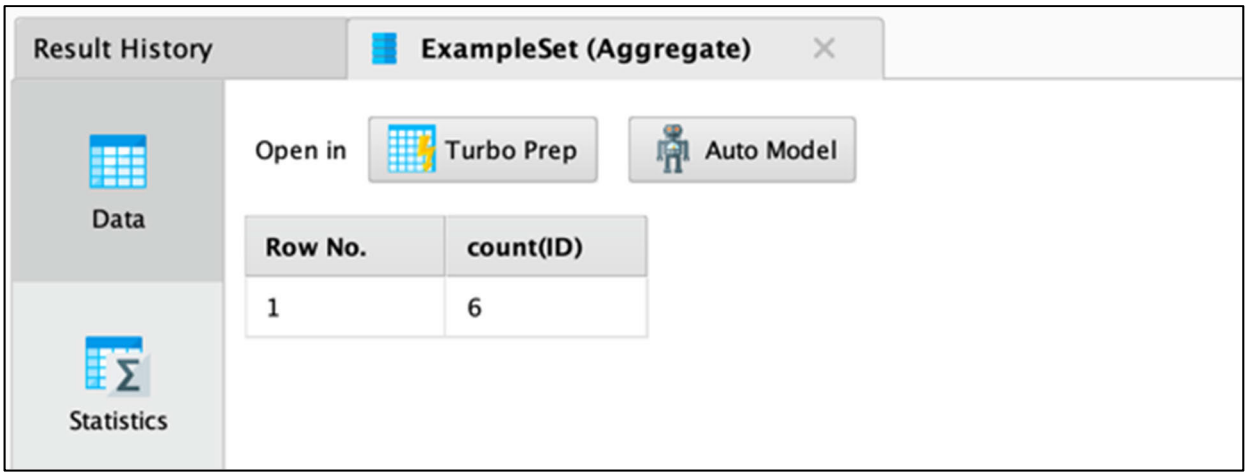

FIGURE 14. The result for the row duplication detection process.

of information. No errors are observed in the aggregated columns. However, the dataset lacked metadata associated with its creation and updates and had no descriptive metadata or standardization for the columns. Moreover, it had no current rows and was published a year after the data availability period.

The Saudi Open Data Portal currently utilizes another significant quality assessment. According to this model, the traffic accident dataset scored 58.33%, as in Table 8. These requirements served as a checklist. If the requirements are met, the value is 1; otherwise, it is 0. The machine readability score was divided among the three requirements according to the open data guidelines published by the portal [47].

To conclude the assessment of the selected dataset, although the scores of the utilized frameworks showed converging results, they scores were 57.8%, 47.74%, 57.3%, and 58.33%, respectively. The shortcomings and strengths of the datasets differ between them.

V. RESULTS AND DISCUSSION

Table 9 shows the quality assessment results of the 15 datasets published in the Saudi Open Data Portal (od.data.gov.sa) using different frameworks. The same frameworks utilized in the use case are used to assess the published open data quality: the framework by Vetrò et al. [37], the proposed model, and the weighted proposed model. The first five rows of table 9 show the results of evaluating the most viewed data at the time this document was written; the middle five rows are the most downloaded datasets and the bottom five rows contain randomly chosen data. The results of the assessment are presented as a score out of 100.

A visual presentation of the assessment is shown in Fig. 6. The frameworks showed some correspondence in fluctuation patterns, but each framework pointed to different quality aspects.

Fig. 7 is a visual representation of the results of the proposed framework, which shows a deficiency in the granularity, publication timeliness, and content timeliness of the examined datasets. This figure shows the need for more information-level completeness and metadata. The unified usage score indicates that most datasets share the same 3-star rating even though the default rating for the datasets is 0, and some have only one view. Therefore, for practical usage detection, publishers should not be allowed to rate their datasets, and users should not be obligated to log in and provide contact information for them to grant the right to rate the datasets.

Fig. 8 shows how the scores in Fig. 7 were affected by the weighing process. Although some criteria experienced a drastic decline in their scores, such as duplication and use, others, such as metadata and granularity, showed a slight change to indicate their significance compared to other criteria.

Fig. 9 shows how the machine readability score indicates that organizations avoid publishing data in a Uniform Resource Identifier (URI) and linked data format. Instead, they presented data in. XLS or.CSV format. The framework by Vetrò et al. [37] shows that the datasets lack traceable updates, current rows, standardized columns, and columns that have metadata. Moreover, the aggregation accuracy score is negatively affected because it does not apply to 8 of the 15 datasets. However, the datasets almost unanimously fulfilled the percentage of syntactically accurate cells. The drastically negative score of −10.4% is attributed to the fact that dataset number 14 had 4170 delayed publication days, which caused a drop in the quality score of the entire dataset.

Fig. 10 shows the evaluation results of the quality assessment model of the portal. It detects a lack of metadata and updated and valuable data. When inspecting the reason behind the low score of requirement number 1.1, the requirements state that ''row data should be provided in Excel or CSV,'' the absence of row data negates the correctness of the sentence. Thus, the fulfillment of this requirement is negated. This dilemma demonstrates the problem of a purely subjective measurement technique. The comprehension of an

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.001">
<context>
<input/>
 <output/>
 <macros/>
     </context>
     <operator activated="true" class="process" compatibility="9.8.001" expanded="true"
name="Process">
       <parameter key="logverbosity" value="init"/>
       <parameter key="random_seed" value="2001"/>
       <parameter key="send_mail" value="never"/>
       <parameter key="notification_email" value=""/>
       <parameter key="process_duration_for_mail" value="30"/>
       <parameter key="encoding" value="SYSTEM"/>
       <process expanded="true">
       <operator activated="true" class="retrieve" compatibility="9.8.001"
expanded="true" height="68" name="Retrieve Test Dataset" width="90" x="112" y="34">
          <parameter key="repository_entry" value="../data/Test Dataset"/>
          </operator>
          <operator activated="true" class="operator_toolbox:tukey_test"
compatibility="2.7.000" expanded="true" height="103" name="Tukey Test" width="90"
x="112" y="187">
            <parameter key="return_preprocessing_model" value="false"/>
            <parameter key="create_view" value="true"/>
            <parameter key="attribute_filter_type" value="all"/>
            <parameter key="attribute" value=""/>
            <parameter key="attributes" value=""/>
            <parameter key="use_except_expression" value="false"/>
            <parameter key="value_type" value="numeric"/>
            <parameter key="use_value_type_exception" value="false"/>
            <parameter key="except_value_type" value="real"/>
            <parameter key="block_type" value="value_series"/>
            <parameter key="use_block_type_exception" value="false"/>
            <parameter key="except_block_type" value="value_series_end"/>
            <parameter key="numeric_condition" value=">=0"/>
            <parameter key="invert_selection" value="false"/>
            <parameter key="include_special_attributes" value="false"/>
          </operator>
          <operator activated="true" class="multiply" compatibility="9.8.001"
expanded="true" height="103" name="Multiply" width="90" x="246" y="85"/>
          <operator activated="true" class="select_attributes" compatibility="9.8.001"
expanded="true" height="82" name="Select Attributes" width="90" x="380" y="238">
            <parameter key="attribute_filter_type" value="single"/>
            <parameter key="attribute" value="TukeyTest_NumberOfOutlierAttributes"/>
            <parameter key="attributes" value=""/>
            <parameter key="use_except_expression" value="false"/>
            <parameter key="value_type" value="attribute_value"/>
            <parameter key="use_value_type_exception" value="false"/>
            <parameter key="except_value_type" value="time"/>
            <parameter key="block_type" value="attribute_block"/>
            <parameter key="use_block_type_exception" value="false"/>
            <parameter key="except_block_type" value="value_matrix_row_start"/>
            <parameter key="invert_selection" value="false"/>
            <parameter key="include_special_attributes" value="false"/>
          </operator>
          <operator activated="true" class="filter_examples" compatibility="9.8.001"
expanded="true" height="103" name="Filter Examples (2)" width="90" x="514" y="187">
            <parameter key="parameter_expression" value=""/>
            <parameter key="condition_class" value="custom_filters"/>
            <parameter key="invert_filter" value="false"/>
            <list key="filters_list">
              <parameter key="filters_entry_key"
value="TukeyTest_NumberOfOutlierAttributes.ge.1"/>
            </list>
            <parameter key="filters_logic_and" value="true"/>
```

Code 1. Accuracy process.

```
            <parameter key="filters_check_metadata" value="true"/>
          </operator>
          <operator activated="true" class="aggregate" compatibility="9.8.001"
expanded="true" height="82" name="Aggregate" width="90" x="648" y="85">
            <parameter key="use_default_aggregation" value="false"/>
            <parameter key="attribute_filter_type" value="all"/>
            <parameter key="attribute" value=""/>
            <parameter key="attributes" value=""/>
            <parameter key="use_except_expression" value="false"/>
            <parameter key="value_type" value="attribute_value"/>
            <parameter key="use_value_type_exception" value="false"/>
           <parameter key="except_value_type" value="time"/>
            <parameter key="block_type" value="attribute_block"/>
            <parameter key="use_block_type_exception" value="false"/>
            <parameter key="except_block_type" value="value_matrix_row_start"/>
            <parameter key="invert_selection" value="false"/>
            <parameter key="include_special_attributes" value="false"/>
            <parameter key="default_aggregation_function" value="average"/>
            <list key="aggregation_attributes">
              <parameter key="TukeyTest_NumberOfOutlierAttributes" value="sum"/>
            </list>
            <parameter key="group_by_attributes" value=""/>
            <parameter key="count_all_combinations" value="false"/>
            <parameter key="only_distinct" value="false"/>
            <parameter key="ignore_missings" value="true"/>
          </operator>
          <connect from_op="Retrieve Test Dataset" from_port="output" to_op="Tukey
Test" to_port="example set input"/>
          <connect from_op="Tukey Test" from_port="example set output" to_op="Multiply"
to_port="input"/>
          <connect from_op="Multiply" from_port="output 1" to_op="Select Attributes"
to_port="example set input"/>
          <connect from_op="Multiply" from_port="output 2" to_port="result 2"/>
          <connect from_op="Select Attributes" from_port="example set output"
to_op="Filter Examples (2)" to_port="example set input"/>
          <connect from_op="Filter Examples (2)" from_port="example set output"
to_op="Aggregate" to_port="example set input"/>
          <connect from_op="Aggregate" from_port="example set output" to_port="result
1"/>
          <portSpacing port="source_input 1" spacing="0"/>
          <portSpacing port="sink_result 1" spacing="0"/>
          <portSpacing port="sink_result 2" spacing="0"/>
          <portSpacing port="sink_result 3" spacing="0"/>
        </process>
      </operator>
    </process>
```

Code 1. ***(Continued.)*** **Accuracy process.**

analyzer may differ from that of a data publisher, because the terminologies that define quality are ambiguous. This finding explains why the critic's viewpoint affected the score.

VI. CONCLUSION

The problem-solving process followed by this study states that, after proposing a solution and generating results, the next step is to evaluate whether the solution confirms the predicted consequences [3]. The proposed framework is expected to detect quality shortcomings that others can not. Table 10 compares the quality issues detected by the frameworks to validate the prediction.

The framework by Vetrò et al. [37] has compatibility limitations with the Saudi Open Data Portal, as the literature review of this study pointed out. The inability to measure the metrics lowered the criterion score because it was assigned zero. As for the consistency of the framework proposed by Vetrò et al. [37], it is believed that the percentage of standardized columns is equivalent to measuring consistency in the framework. The ratio of standardized columns is calculated to ''the number of columns that represent some kind of information that has a standard associated with it (i.e., geographic information)'' [37] since the Saudi open data portal does not enforce standardization on the data, this affected the consistency score in the framework. The proposed framework measures data consistency differently by depicting the number of cells that did not match the column format, which was not an issue for most of the Saudi dataset. The difference in

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.001">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.001"
expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.001"
expanded="true" height="68" name="Retrieve Test Dataset" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../data/Test Dataset"/>
      </operator>
      <operator activated="true" class="remove_duplicates" compatibility="9.8.001"
expanded="true" height="103" name="Remove Duplicates" width="90" x="246" y="34">
        <parameter key="attribute_filter_type" value="all"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
        <parameter key="treat_missing_values_as_duplicates" value="false"/>
      </operator>
      <operator activated="true" class="aggregate" compatibility="9.8.001"
expanded="true" height="82" name="Aggregate" width="90" x="380" y="136">
        <parameter key="use_default_aggregation" value="false"/>
        <parameter key="attribute_filter_type" value="all"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
        <parameter key="default_aggregation_function" value="average"/>
        <list key="aggregation_attributes">
          <parameter key="_id" value="count"/>
        </list>
        <parameter key="group_by_attributes" value=""/>
        <parameter key="count_all_combinations" value="false"/>
        <parameter key="only_distinct" value="false"/>
        <parameter key="ignore_missings" value="true"/>
      </operator>
      <connect from_op="Retrieve Test Dataset" from_port="output" to_op="Remove
Duplicates" to_port="example set input"/>
      <connect from_op="Remove Duplicates" from_port="duplicates" to_op="Aggregate"
to_port="example set input"/>
      <connect from_op="Aggregate" from_port="example set output" to_port="result
1"/>
```

Code 2. Row duplication process.

```
            <portSpacing port="source_input 1" spacing="0"/>
            <portSpacing port="sink_result 1" spacing="0"/>
            <portSpacing port="sink_result 2" spacing="0"/>
          </process>
        </operator>
      </process>
```

Code 2. ***(Continued.)*Row duplication process.**

defining consistency indicates that the proposed framework targeted what the Saudi Open Data Portal enforces in its practices. Moreover, the proposed framework cannot measure ''data quality'' and ''valuable data'' because it is a subjective assessment measure. The portal's lack of clear indicators of what is considered valuable and high quality affects the score of the two metrics by lowering it, making the readings unreliable and thus incompatible with the nature of the Saudi portal.

The comparative findings demonstrated that the intended effect of the suggested solution was successfully attained. The proposed framework is demonstrated to cover the gap left by the current quality evaluation frameworks to address the issue of undetected quality inadequacy, while simultaneously targeting the quality metrics that reflect the operational norms of the Saudi open data portal.

Through investigating towards the research findings, the following questions are answered:

RQ1. What is the impact that the Saudi open government data portal made so far?

The literature review showed that the Saudi Open Data Portal did not achieve its intended goals. The portal has a low impact on social indicators, international benchmarking, and publication and research on the open data in Saudi Arabia.

RQ2. How can the portal's performance be improved to assist it in achieving its objectives?

Section II-C linked the mediocre impact of the portal with the poor data quality that feeds it. This section also argues that constant monitoring and evaluation are key to improving performance. Thus, this study attempted to construct a customized quality assessment framework that addresses the needs of the Saudi open data portal and accurately assesses its quality status.

RQ3. What benefits can come from investment in the open data infrastructure?

The last section of this study, titled 'Recommendations provides suggestions for changes in the OGD management process. It also includes some possible benefits of investing in OGD improvement.

A. RESEARCH LIMITATIONS

The proposed framework remains a simple model with the potential for enhancement after further performance evaluation. The limitations of this study are as follow:

1. Gathering information to calculate the ILC of the proposed framework is challenging as an outsider to a government organization in Saudi Arabia, which restricts the ability to evaluate more samples.
2. In the Results and Discussion section, the quality assessment scores for the framework by Vetrò et al. [37] and the open-data portal framework was mostly subjective. Thus, the assessed datasets were given the benefit of doubt by scoring the measures as much as possible to protect the integrity of the study.
3. The proposed framework keeps altering upon further testing, implying that further adjustments may be required.
4. The ILC criterion cannot detect missing dataset records. An example is a dataset that publishes all non-private attributes but has only five rows out of the database's hundreds of records. The framework will not be able to detect an issue with completeness in this case.
5. The study did not analyze the perspective of OGD publishers, which are Saudi government agencies. Knowing the data providers' challenges and limitations of data providers can help examine the issue from all angles and create a comprehensive solution. However, due to time restrictions and acceptance limitations that was not possible.

B. FUTURE WORK

Based on the limitations of this study, further testing on portal datasets is recommended. Furthermore, the proposed framework should be constantly enhanced based on updates in the best practices and feedback and results from expanding the testing process. The economic impact of OGD on Saudi Arabia is an important topic that requires further research to track the progress of its initiatives and towards its goals.

VII. RECOMMENDATIONS

This study provides recommendations based on the experience of navigating the portal to assess the quality of its datasets. Starting with suggestions to improve the user experience and listing the benefits the Saudi government may obtain from maximizing the potential of its OGD infrastructure.

A. FEEDBACK AND SUGGESTIONS FOR THE SAUDI OPEN DATA PORTAL

This section lists some notes for Saudi OGD stakeholders in an attempt to provide feedback to help improve the portal's user experience.

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.001">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.001"
expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.001"
expanded="true" height="68" name="Retrieve Test Dataset" width="90" x="45" y="85">
        <parameter key="repository_entry" value="../data/Test Dataset"/>
      </operator>
      <operator activated="true" class="transpose" compatibility="9.8.001"
expanded="true" height="82" name="Transpose" width="90" x="179" y="136"/>
      <operator activated="true" class="remove_duplicates" compatibility="9.8.001"
expanded="true" height="103" name="Remove Duplicates" width="90" x="313" y="136">
        <parameter key="attribute_filter_type" value="all"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
        <parameter key="treat_missing_values_as_duplicates" value="false"/>
      </operator>
      <operator activated="true" class="aggregate" compatibility="9.8.001"
expanded="true" height="82" name="Aggregate" width="90" x="581" y="238">
        <parameter key="use_default_aggregation" value="false"/>
        <parameter key="attribute_filter_type" value="all"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
        <parameter key="default_aggregation_function" value="average"/>
        <list key="aggregation_attributes">
          <parameter key="id" value="count"/>
        </list>
        <parameter key="group_by_attributes" value=""/>
        <parameter key="count_all_combinations" value="false"/>
        <parameter key="only_distinct" value="false"/>
        <parameter key="ignore_missings" value="true"/>
      </operator>
      <operator activated="true" class="transpose" compatibility="9.8.001"
expanded="true" height="82" name="Transpose (2)" width="90" x="447" y="136"/>
```

Code 3. Column duplication process.

```
          <operator activated="true" class="select_attributes" compatibility="9.8.001"
expanded="true" height="82" name="Select Attributes" width="90" x="581" y="136">
            <parameter key="attribute_filter_type" value="single"/>
            <parameter key="attribute" value="id"/>
            <parameter key="attributes" value="id"/>
            <parameter key="regular_expression" value="id"/>
            <parameter key="use_except_expression" value="false"/>
            <parameter key="value_type" value="attribute_value"/>
            <parameter key="use_value_type_exception" value="false"/>
            <parameter key="except_value_type" value="time"/>
            <parameter key="block_type" value="attribute_block"/>
            <parameter key="use_block_type_exception" value="false"/>
            <parameter key="except_block_type" value="value_matrix_row_start"/>
            <parameter key="invert_selection" value="true"/>
            <parameter key="include_special_attributes" value="true"/>
          </operator>
          <connect from_op="Retrieve Test Dataset" from_port="output" to_op="Transpose"
to_port="example set input"/>
          <connect from_op="Transpose" from_port="example set output" to_op="Remove
Duplicates" to_port="example set input"/>
          <connect from_op="Remove Duplicates" from_port="example set output"
to_op="Transpose (2)" to_port="example set input"/>
          <connect from_op="Remove Duplicates" from_port="duplicates" to_op="Aggregate"
to_port="example set input"/>
          <connect from_op="Aggregate" from_port="example set output" to_port="result
2"/>
          <connect from_op="Transpose (2)" from_port="example set output" to_op="Select
Attributes" to_port="example set input"/>
          <connect from_op="Select Attributes" from_port="example set output"
to_port="result 1"/>
          <portSpacing port="source_input 1" spacing="0"/>
          <portSpacing port="sink_result 1" spacing="0"/>
          <portSpacing port="sink_result 2" spacing="0"/>
          <portSpacing port="sink_result 3" spacing="0"/>
        </process>
      </operator>
    </process>
```

Code 3. ***(Continued.)*** **Column duplication process.**

1) THE PORTAL'S MAINTAINERS AND DEVELOPERS

1. Searching the portal is challenging. Users are expected to enter keywords from the dataset's title only when, in reality, they could be searching for a dataset that is published on a specific date. Unfortunately, the website did not address this issue. Searching for a dataset within the publishers' list is even more restricted, because the user is required to enter the exact title. The search bar does not recognize a keyword or even part of the title, which can be frustrating for the user. Dual-language searching is also a challenge; the user receives different results when searching for the same topic in English and Arabic. This gap can make the discovery of information difficult for monolinguals.
2. Some publishers had an empty list of publications. Encouraging publishers to use the portal instead of publishing the data only on their official websites will help people find all the data they seek from one resource. As for reticent organizations, training and guidance will help them understand the portal's vision, and their national duty may encourage them to participate in the portal. One suggestion to encourage participation is to have an annual event organized by the OGD monitoring authority to reward active organizations that publish valuable data for re-users.
3. Hosting seminars and workshops for government agencies can streamline their efforts and spread awareness among publishers regarding what is expected and how they can achieve it. These events can also serve as opportunities for participants to exchange experiences.
4. The portal should have a strict quality measurement system. While efforts to audit the published data are appreciated, according to the data quality guidelines, the quality assurance process must ensure that the data are reusable not just informative.
5. Incorporating active visual representations such as maps and charts is essential. Citizens who visit the portal come from different academic backgrounds and not all can interpret spreadsheets. Furthermore, clearly

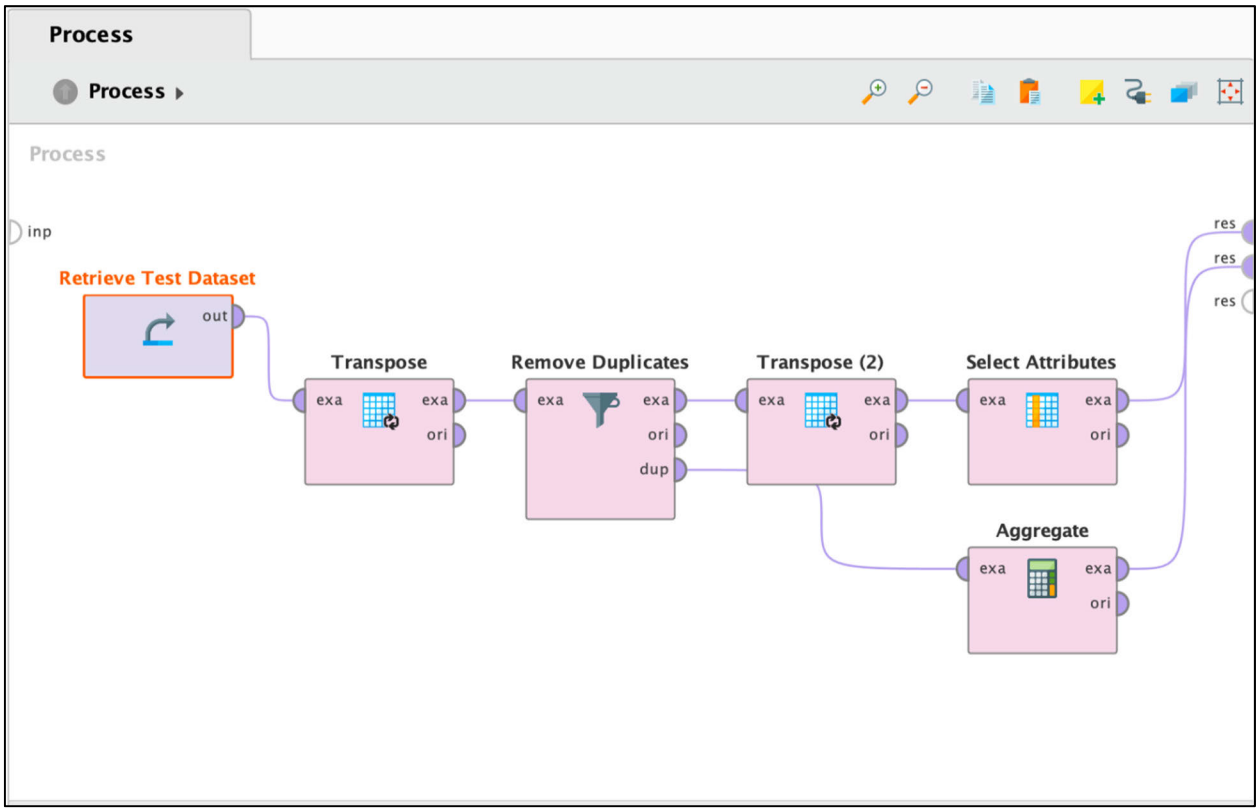

FIGURE 15. Column duplication process in RapidMiner.

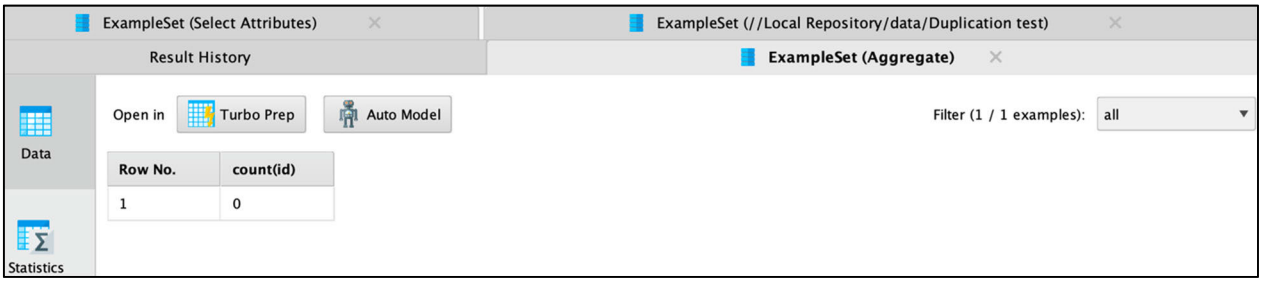

FIGURE 16. The results for the column duplication process.

presenting the data can help a broader segment of people utilize it.

6. The re-user should be able to reference a dataset with confidence that its data will not change after using it. Data currency requires primary forms of change data capture (CDC) for data warehousing. The CDC can be enforced in the open data portal by: a) providing updates on the dataset in the form of versions while maintaining the previous version in its original state and prohibiting edits. Allowing the re-user to reference the version they utilized with the assurance that the data will not change; or b) the dataset can be given a timestamp column that indicates the last update provided in the rows, allowing for easy calculation of the number of edited rows. None of the CDC practices are applied to the Saudi open data portal. Furthermore, according to the Saudi portal's data quality guidelines, publishers must update the dataset by adding new data to the same file without using any differentiation measures for the latest data. This can affect the integrity of data over time.
7. Linking a "related dataset" should be an automatic feature instead of a manual feature. Not all publishers took the time to link their datasets. Therefore, automation can be performed using technologies, such as artificial intelligence and linked data.
8. Minor details, such as metadata, data dictionaries, maintainers' contact details, and dataset size, can drastically improve user experience. Thus, the portal must ensure that the publisher includes descriptive information before publication.
9. Not all content is correctly translated from Arabic to English, and vice versa. The developers must assume the portal does not cater only to the data community

TABLE 12. Metadata checklist.

ID	Metadata Identifier	Description	example	Weight	Availability (1= available / 0 = not available)
1	Dataset Title	The title of the resource (dataset, document, picture … etc.)	"Salaries Scales for Civil Servants"	2	☐
2	Dataset description	A brief and accurate description of the information that the resource hold	"the salaries of the government employees based on their positions, status and employment agencies"	2	☐
3	Dual-language	The available title is provided in Arabic and English to ensure accessibility	العنوان: "سلم رواتب موظفي الدولة في قطاع الخدمة المدنية"	1	☐
4	categorization	The resource should be assigned to at least one of the predefined categories.	Finance, health, education … etc.	2	☐
5	Keywords	Assigning keywords to the resource help with categorization, search and reference	Accounts Financial Monetary Affairs and Industry, Labor Market, Open Data Migration Group, Prices and Indices	2	☐
6	publisher name	The name of the organization responsible for publishing the resource.	Ministry of Civil Service	1	☐
7	publisher Link	A URL to the publishers' main website or the publishers' main portal.	https://www.mcs.gov.sa/en/Pages/default.aspx	1	☐
8	URL for the dataset	The resource should have a sharable direct URL to it, taking into account different format of the resource.	https://data.gov.sa/Data/ar/dataset/47b30d64-4382-48a3-ac7a-0b5124e9c57d/resource/bc891542-8c2c-4179-abf5-b4b1d7a7d69d/download/salaries_scales.csv	1	☐
9	format	List the file types available for the resource	cvs, xls	1	☐
10	Resource size	The size of the resource in byte, according to the quality measures of the Saudi portal it should not exceed 20 MB	430 KB	1	☐
11	Maintainer	The name of the department or the person responsible for the published content.	Ministry of Civil Service – 9276 Prince Saad Ibn Abdul Aziz, AL-Namothajiah district ،Riyadh 12731 4731	1	☐
12	Maintainer contact	E-mail, phone number or social media identifiers of the maintainer.	customerservice@mcs.gov.sa	2	☐
13	Publishing or creation date	The date that the resource made available.	May 2, 2019, 11:01 (AST)	2	☐
14	Last modified on	The date of the last update.	December 24, 2019, 09:40 (AST)	2	☐

TABLE 12. *(Continued.)* **Metadata checklist.**

15	Update frequency	How often is the resource set for an update?	Quarterly	2	☐
16	Data dictionary	Human-readable description of the data fields in the dataset	Organization: the name of the government agency; Department: the name of the department that the position belongs to; Salary: the amount in SAR	1	☐
			Total		

elite. Not all local users can understand English; however, the portal includes datasets that are only available in English. On the other hand, international audiences unfamiliar with the Arabic language can encounter many datasets available only in Arabic. The Spanish OGD portal also faced the same issues. It took much time to understand the content, even with a poorly translated website. It seemed that it was designed to cater only to a Spanish-speaking audience, which contradicts the purpose of having open data in the first place.

10. Publishers were requested to upload datasets of less than 20 MB in size. However, the portal should refrain from publishing big data, as researchers and data analysts find it exciting and valuable. The dataset size should be listed in the ''additional info'' box to inform users of the size before they decide to download it.
11. Activating connection channels, including e-mail addresses, live chat services, and phone numbers, is crucial when managing an international portal. The Saudi Food and Drug Authority (SFDA) has established an outstanding model with its public contact service. The portal can benefit from experience and replicate it. Furthermore, it is highly recommended to ease the restriction on the user's interaction, as the user must provide a considerable amount of information to leave a comment or inquiry.

2) PUBLISHERS

1. Publishers are advised to focus on the timeliness of their data. Publishing archived reports from decades ago will not help the public without recent data to form a comprehensive perspective. Certain publishers frequently update their archives without changing the content of datasets, which is pointless and time-consuming.
2. Organizations should avoid publishing summarized reports, and re-users need raw data. Instead of processing data, organizations should publish non-private attributes from their databases. The re-users can then use them to generate reports. Specific attributes may seem insignificant for publishing, but they can help others. For instance, reports on duties collected on imports state the yearly income of Saudi Arabia in Riyals. Here, it would be better to provide a detailed description of each imported product with its value, type of product, source country, and destination city. Complete datasets will help entrepreneurs understand the gap between what consumers need and what local stores are not providing. Another example is accident reports. Instead of showing the number of accidents that occur annually, a dataset that details each accident, including the road's name, details about the cars, and weather conditions, could help identify patterns and mitigate the causes of accidents.
3. Participation in an OGD initiative as a government agency is challenging. Regular data selection, auditing, quality assurance, publication, and maintenance are required. It also requires constant collaboration within the organization's departments and between the organization and the public. To manage these tasks effectively, each publisher can assign a designated OGD team with data specialists to manage open data publications.
4. Data that seems easy to interpret within an organization may need to be clarified for people outside the organization. Explaining terminologies and describing datasets and metadata can help re-users avoid misinterpretation.
5. Publishers must also consider the target audience. People from different backgrounds browse the portal. The average user knowledge may include data specialists, inexperienced individuals, or potential international investors. Thus, it is vital to consider how all of them perceive published data.
6. It is best to provide a link to the agency's web page on the portal instead of publishing data on the official agency's website. This saves time and avoids maintenance of duplicate versions of datasets.

3) DATA ANALYTICS COMMUNITY

1. Saudi data communities are advised to keep requesting the data they need from a specific organization and the portal. Even if their requests are not met, it gives the OGD presenters an idea of their audiences' demands.
2. Participation by leaving a comment and rating the open datasets is critical for improving open data performance. Feedback will help to create a user-oriented experience instead of broadcasting a government-oriented platform.

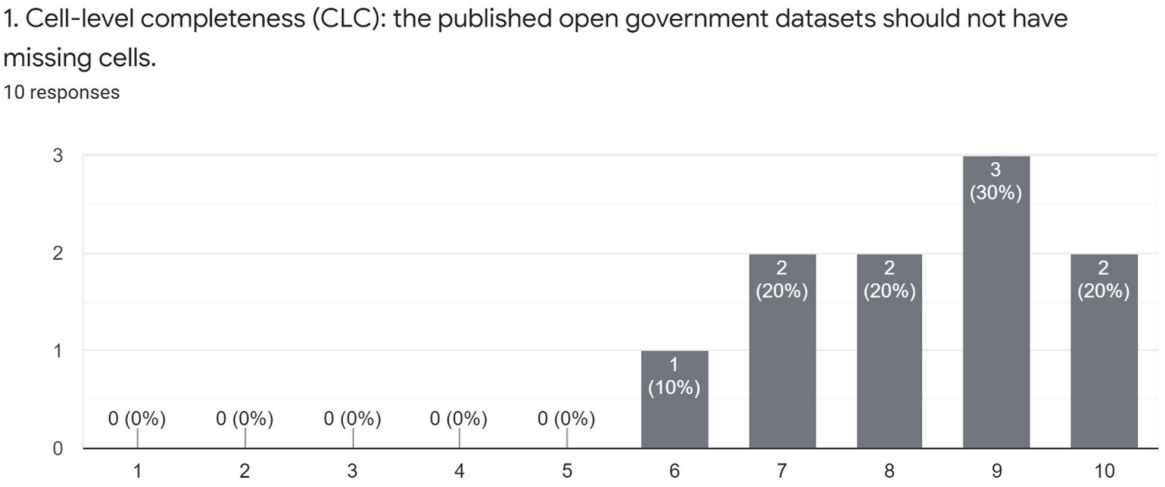

FIGURE 17. Survey results for cell level completeness.

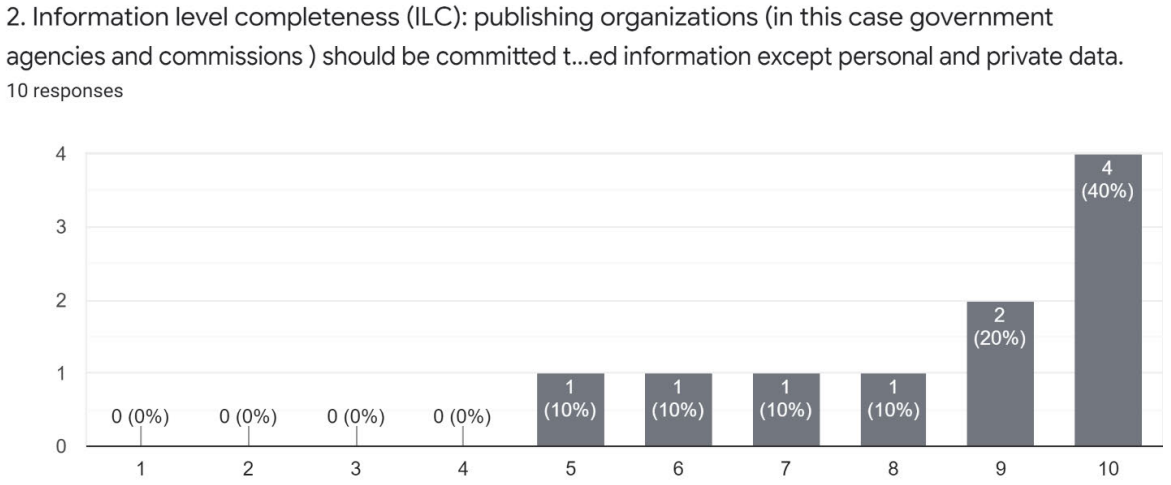

FIGURE 18. Survey results for information level completeness.

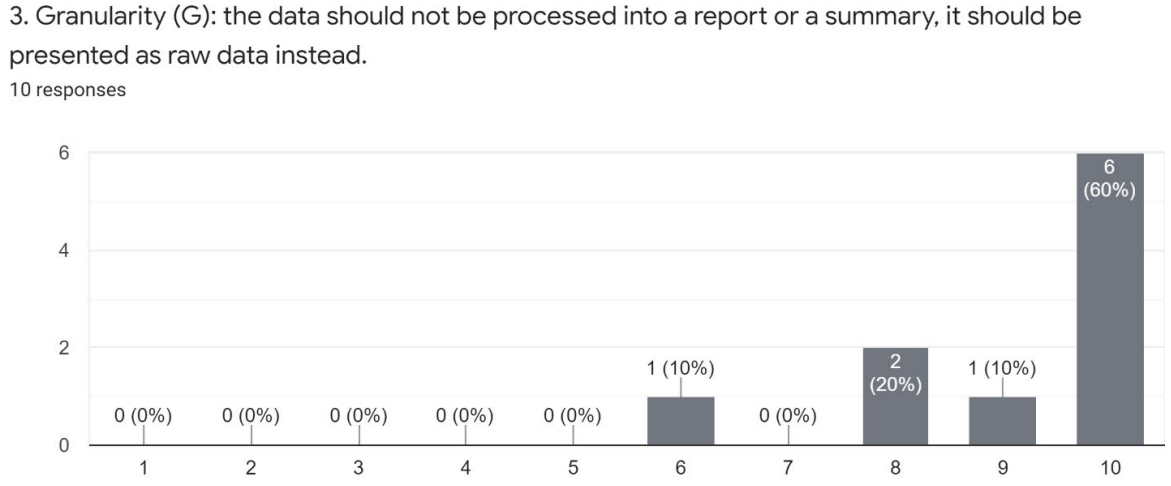

FIGURE 19. Survey results for granularity.

B. HOW CAN SAUDI ARABIA BENEFIT FROM AN EFFECTIVE OGD?

Inspired by McKinsey's report on open data [17], this section discusses the effect that proper implementation of the OGD approach can have on Saudi Arabia: 1) *Education*: Open performance data in universities can help attract investors, and in schools, these data can help trace their productivity and provide more jobs for teachers using accurate job vacancy

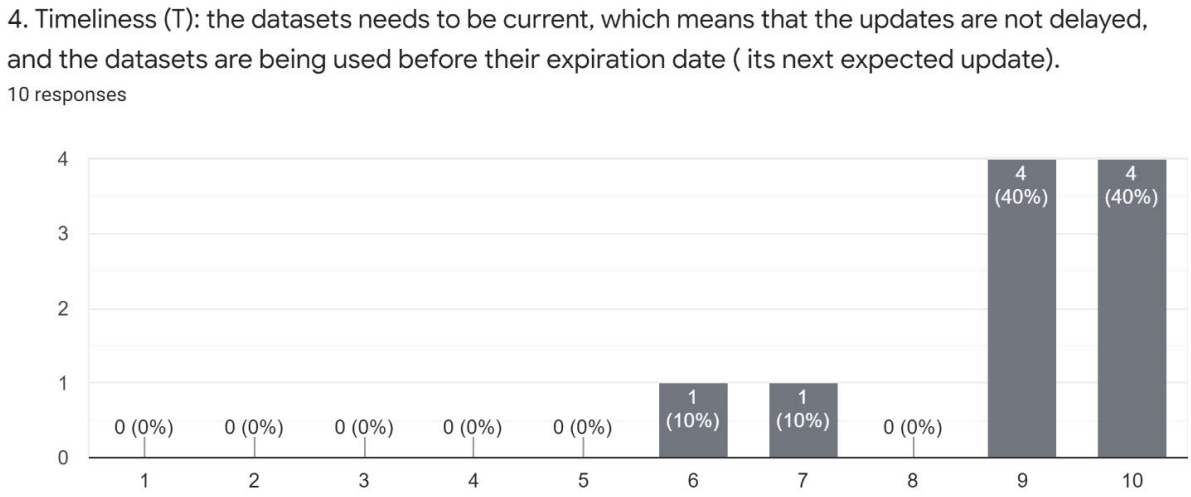

FIGURE 20. Survey results for timeliness.

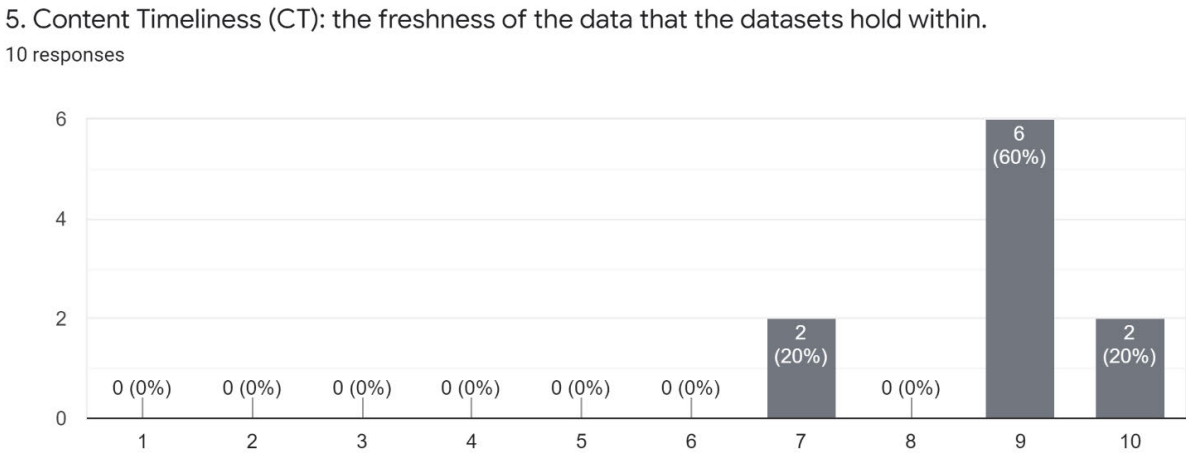

FIGURE 21. Survey results for content timeliness.

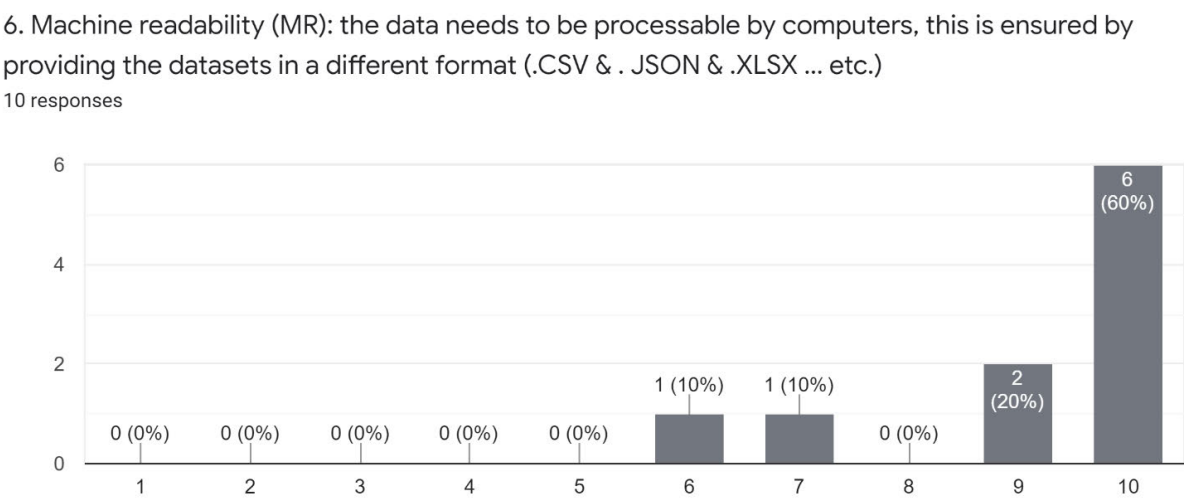

FIGURE 22. Survey results for machine readability.

detection. In general, the Ministry of Education will be able to perform better resource allocation, enhanced decision-making, and optimization of strategies. 2) *Transportation*: Access to real-time location and traffic data can reduce travel times by identifying alternative routes to destinations to mitigate traffic, adjust public transportation schedules to match public demand, and achieve better transportation management and long-term positive environmental impacts. 3) *Trad-*

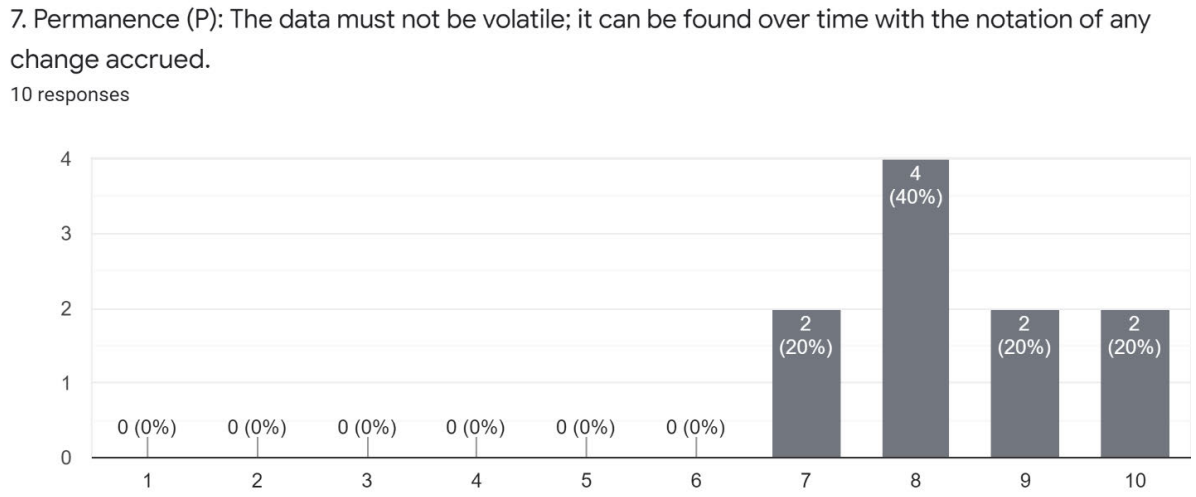

FIGURE 23. Survey results for permanence.

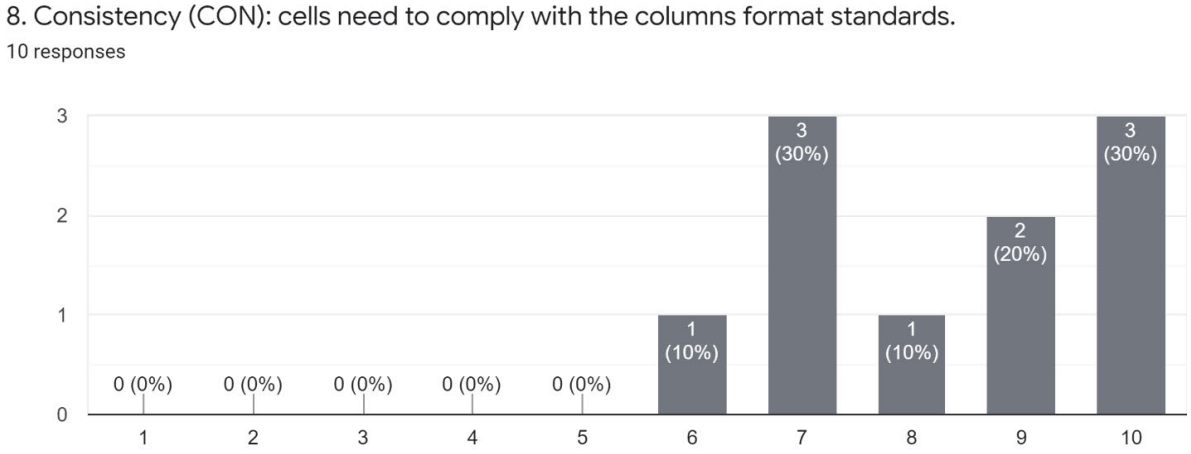

FIGURE 24. Survey results for consistency.

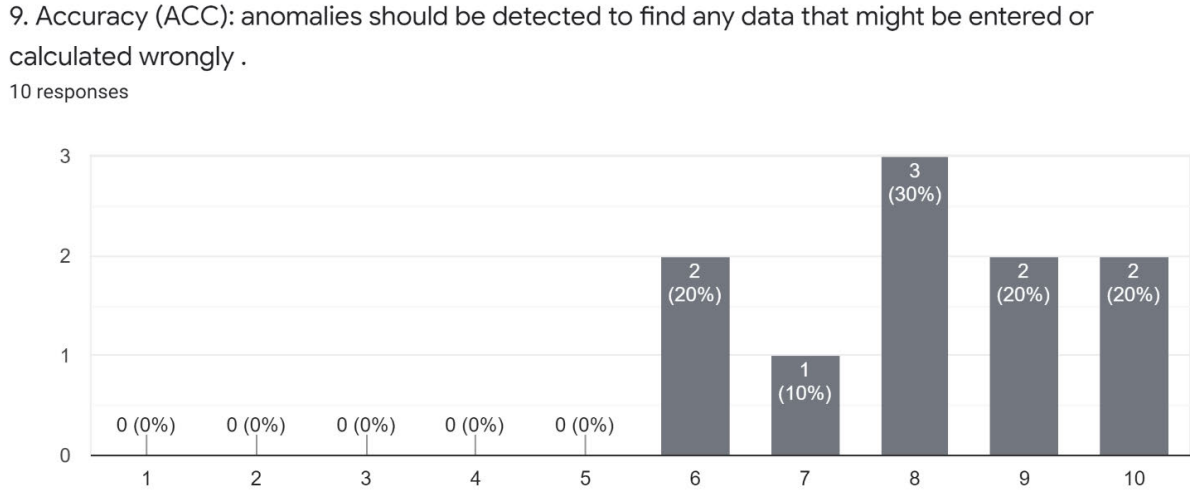

FIGURE 25. Survey results for accuracy.

ing: Utilizing data about demographics and store assortment for a specific neighborhood can help businesses meet the needs of their community, segment consumers, and customize consumer services, products, and strategies in the Saudi market. 4) *Electricity*: Provides citizens with detailed data about their energy consumption and shows how similar houses or

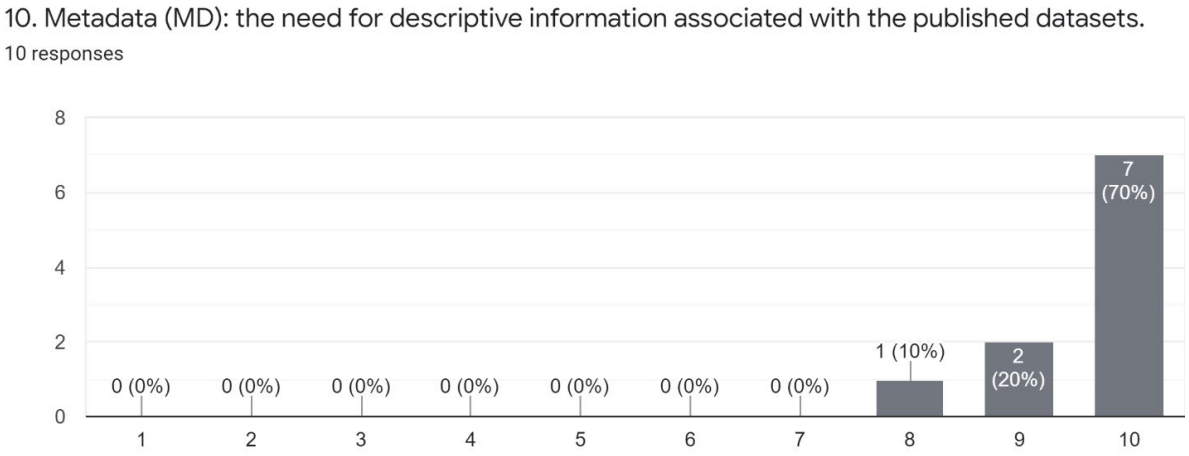

FIGURE 26. Survey results for metadata.

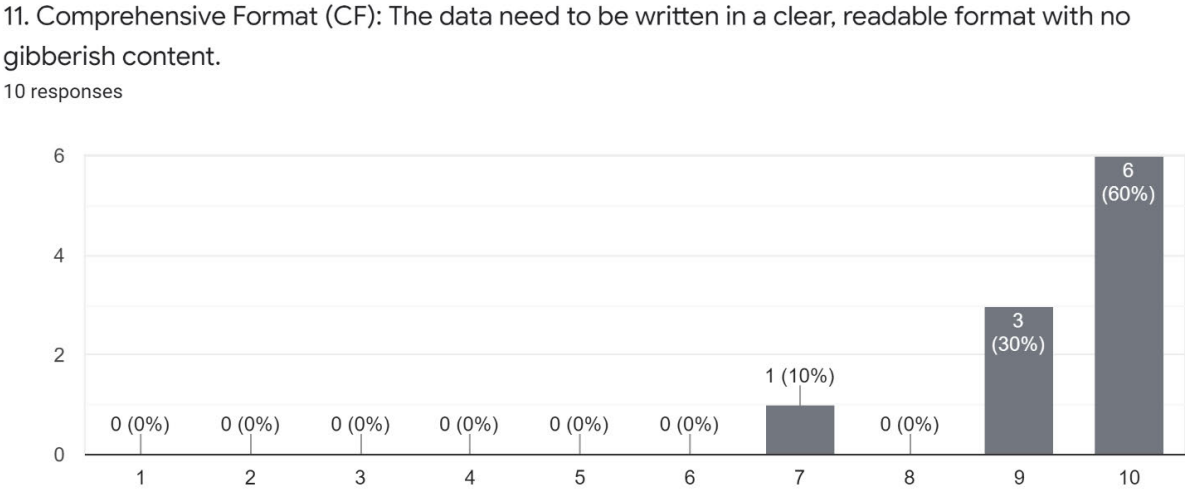

FIGURE 27. Survey results for comprehensive format.

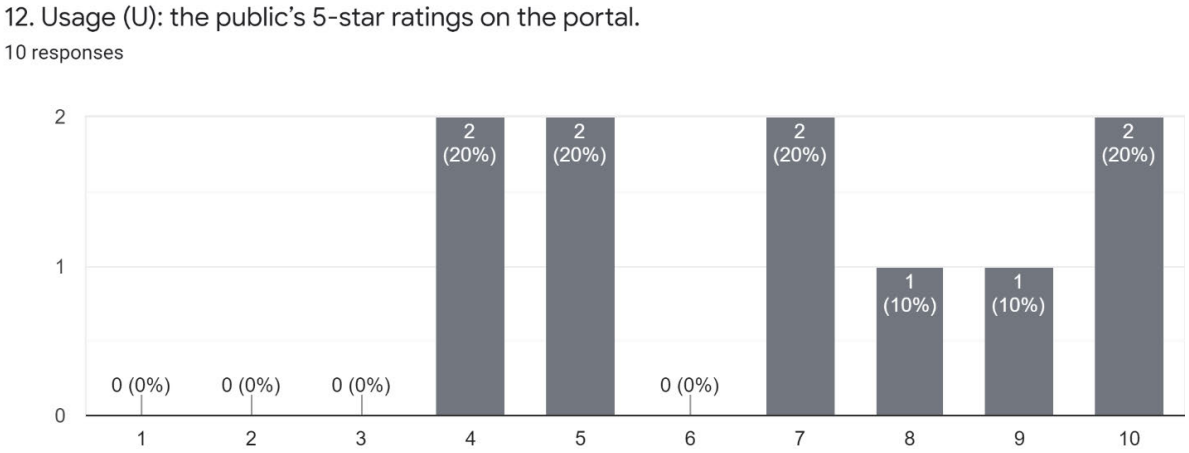

FIGURE 28. Survey results for usage.

businesses use electricity. High-level statistics for spreading awareness of energy regulation are not as effective as providing localized and personalized statistics and benchmarks. 5) *Oil and Gas*: Sharing consumption data helps citizens make better-informed decisions about energy use and helps oil and gas companies achieve better allocation of their facilities; 6) *Health care*: helps patients allocate the most timely and appropriate treatment, better emergency response, and better resource management and campaign impacts for the Ministry of Health. 7) *Housing*: Open data can help match buyers

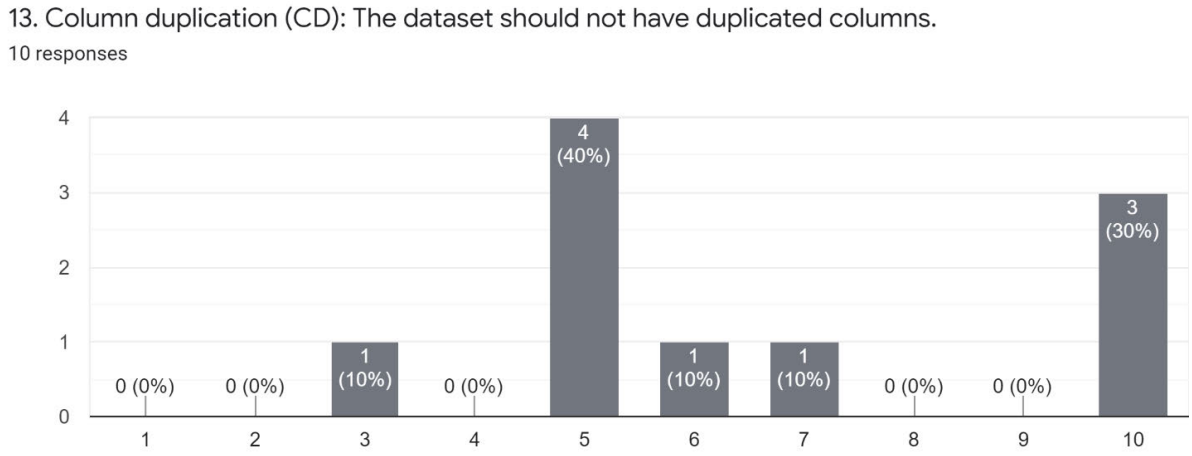

FIGURE 29. Survey results for column duplication.

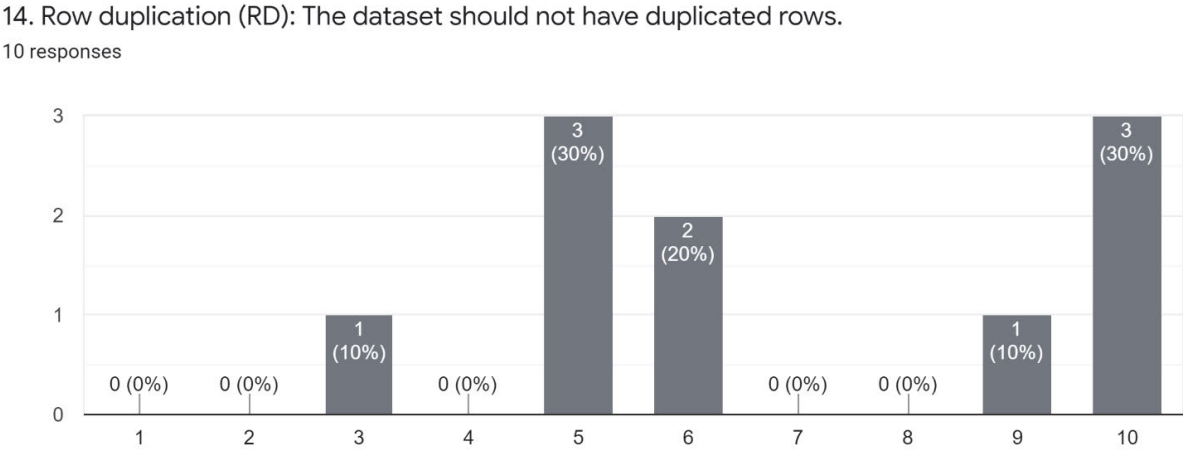

FIGURE 30. Survey results for row duplication.

and renters with suitable properties and help municipalities optimize the layout of the infrastructure in the development of neighborhoods.

APPENDIX A
OPEN GOVERNMENT DATA QUALITY METRICS

See Table 11.

APPENDIX B
RAPIDMINER PROCESSES

C. ACCURACY

XML code 1 of the accuracy process demonstrated in Fig. 11 performs the Tukey test on the dataset and then counts the number of outlier cells that are marked as ''top outlier'' or ''bottom outlier'' to give the total number of outliers, as shown in Fig. 12.

D. REDUNDANCY

1) ROW DUPLICATION (NDR)

To find the number of duplicated rows the process shown in XML Code 2 and Fig. 13 uses ''Remove Duplicates'' operator, the filtered rows are counted using ''Aggregate'' operator after setting the aggregation attribute to the primary key of the dataset, this arrangements will present NDR as shown in Fig. 14.

2) COLUMN DUPLICATION (NDC)

Fig. 15 illustrates the XML code 3 process for detecting column duplication, which utilizes the 'transpose' operator that shifts columns to rows. For this process, the dataset is initially transposed to make the columns into rows, this will allow the operator ''remove duplicates'' to filter out duplicated rows before returning them as columns. Counting the duplicates using the ''Aggregation'' operator will result in the NDC, as shown in Fig. 16.

APPENDIX C
METADATA CHECKLIST

See Table 12.

APPENDIX D
SURVEY RESULTS

E. SURVEY DESCRIPTION

1) OPEN GOVERNMENT DATA QUALITY IN SAUDI ARABIA

In this survey, we aimed to measure the importance of open government data quality characteristics that matter most to

the open data community and the data quality community in Saudi Arabia.

Please rate the following 14 quality characteristics based on their importance according to your assessment (with 1 indicating that the characteristic is not important to the quality of the open government dataset and 10 indicating that the characteristic is crucial to the quality of the datasets).

F. PARTICIPANT'S QUALIFICATIONS

Participant NO.1: Associate Professor at King Saud University, pioneer in the field of data science.

Participant NO.2: A Director of Analysis and Performance Measurement Department.

Participant NO.3: Co-founder of Lucidia and a big data specialist.

Participant NO.4: A Specialist in data science and R language, visiting researcher at the University of Leeds, and PhD in applied statistics and big data.

Participant NO.5: Data scientist and statistician, and a consultant on data analysis.

Participant NO.6: General supervisor of the Big Data Center in the Emirate of Makkah.

Participant NO.7: Assistant professor at Al-Jouf University, specializing in data science and big data.

Participant NO.8: Associate Professor at Imam Muhammad bin Saud Islamic University, specializing in open data, data analysis, and management.

Participant NO.9: Data Scientist and Artificial Intelligence Consultant at Oracle Middle East, and founder of Bayan Enriching Data Science Platform.

Participant NO.10: General Director of Data at the Real Estate General Authority and has a Master's degree in Big Data.

G. SURVEY RESULTS

See Figures 17–30.

REFERENCES

[1] Digital Government Authority. *Open Government Data*. Accessed: Mar. 22, 2023. [Online]. Available: https://www.my.gov.sa/wps/portal/snp/eParticipation/openData/!ut/p/z1/jZDbCoJAEEC_xldnNrWW3jYzUrItutm-hMG2CubGZvn7iUFQdJu3Gc6Bw4CABESZXnOVVrku06LZt6K7i0M_RO4Sjt7awTkLpq4bDTvYc2DTAtGMuoQh4dzxBjj34x5nizVB9ED84-OHYfjbFy3yKKA0DJqCMR35yxiRk1fgTWILfGlYpAYiEKrQ-

[2] E. Huyer, L. V. Knippenberg, and E. L. Arriens, "The economic impact of open data—Opportunities for value creation in Europe," European Data Portal, European Commission, Tech. Rep., 2020.

[3] J. L. Kolodner, R. L. Simpson, and K. Sycara-Cyranski, "A process model of cased-based reasoning in problem solving," in *Proc. 9th Int. Joint Conf. Artif. Intell.*, Los Angeles, CA, USA, 1985, pp. 284–290.

[4] R. K. Merton, "The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property," *Isis*, vol. 79, no. 4, pp. 606–623, Dec. 1988, doi: 10.1086/354848.

[5] D. Berliner, A. Ingrams, and S. J. Piotrowski, "The future of FOIA in an open government world: Implications of the open government agenda for freedom of information policy and implementation," *Villanova Law Rev.*, vol. 63, pp. 867–894, Dec. 2018.

[6] A. B. Bode. *Open Data: A History, Data.Gov*. Accessed: Feb. 22, 2020. [Online]. Available: https://www.data.gov/blog/open-data-history

[7] *The Annotated 8 Principles of Open Government Data*. Accessed: Apr. 5, 2023. [Online]. Available: https://opengovdata.org/

[8] The White House. *Open Goverment Initiative*. Accessed: Mar. 10, 2019. [Online]. Available: https://obamawhitehouse.archives.gov/open

[9] Open Government Partnership. *About Open Government Partnership*. Accessed: Oct. 30, 2019. [Online]. Available: https://www.opengovpartnership.org/about

[10] United States Congress. *Digital Accountability and Transparency Act of 2014*. Accessed: Feb. 1, 2020. [Online]. Available: https://www.congress.gov/113/plaws/publ101/PLAW-113publ101.pdf

[11] Å. Grönlund and T. A. Horan, "Introducing e-gov: History, definitions, and issues," *Commun. Assoc. Inf. Syst.*, vol. 15, pp. 713–729, Jan. 2005, doi: 10.17705/1CAIS.01539.

[12] B. Egidijus, *Exploring Digital Government Transformation in the EU*. Publications Office of the European Union, 2019.

[13] SDAIA. (Mar. 2023). *OPEN DATA HANDBOOK*. Accessed: May 7, 2023. [Online]. Available: https://od.data.gov.sa/downloads/booklet/guidelines/Open%20Data%20Handbook_En.pdf

[14] General Authority for Statistics. *The Total Population, The General Authority for Statistics*. Accessed: Mar. 29, 2023. [Online]. Available: https://www.stats.gov.sa/en/node

[15] A. Zuiderwijk, A. Pirannejad, and I. Susha, "Comparing open data benchmarks: Which metrics and methodologies determine countries' positions in the ranking lists?" *Telematics Informat.*, vol. 62, Sep. 2021, Art. no. 101634, doi: 10.1016/j.tele.2021.101634.

[16] *World Wide Web Foundation, Open Data Barometer—Leaders Edition*, World Wide Web Foundation, Washington, DC, USA, 2018.

[17] Open Knowledge. *Global Open Data Index—Methodology*. Accessed: Mar. 30, 2023. [Online]. Available: http://2015.index.okfn.org/methodology/

[18] J. Henninger, E. Swanson, L. Noe, T. Wahabzada, A. Pittman, and T. Hadnot, "ODIN open data inverntory biennial report 2022/23," Tech. Rep., 2022.

[19] *E-government Survey 2022 the Future of Digital Government*, United Nations, New York, NY, USA, 2022.

[20] Z. S. Alzamil and M. A. Vasarhelyi, "A new model for effective and efficient open government data," *Int. J. Discl. Governance*, vol. 16, no. 4, pp. 174–187, Aug. 2019, doi: 10.1057/s41310-019-00066-w.

[21] E. M. Asyri and M. D. Alsuraihi, "The open data platform and its activation through the e-government portals of the Gulf Cooperation Council countries: A comparative study," *J. Inf. Stud. Technol.*, vol. 2018, no. 2, 2019.

[22] S. Saxena, "National open data frames across Japan, The Netherlands and Saudi Arabia: Role of culture," *Foresight*, vol. 20, no. 1, pp. 123–134, Mar. 2018, doi: 10.1108/fs-07-2017-0038.

[23] Hofstede Insights. *Country Comparison Tool*. Accessed: Oct. 31, 2019. [Online]. Available: https://www.hofstede-insights.com/country-comparison/

[24] S. Saxena, "Evaluation of the national open government data (OGD) portal of Saudi Arabia," in *Politics and Technology in the Post-Truth Era*, A. Visvizi and M. D. Lytras, Eds. Bingley, U.K.: Emerald Publishing Limited, 2019, pp. 221–235.

[25] R. Máchová, M. Hub, and M. Lnenicka, "Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective," *Aslib J. Inf. Manag.*, vol. 70, no. 3, pp. 252–268, May 2018, doi: 10.1108/ajim-02-2018-0026.

[26] M. W. and A. Khader, "Measuring the data openness for the open data in Saudi Arabia e-government—A case study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 12, pp. 113–122, 2016, doi: 10.14569/ijacsa.2016.071215.

[27] A. Abella, M. Ortiz-De-Urbina-Criado, and C. De-Pablos-Heredero, "The process of open data publication and reuse," *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 3, pp. 296–300, Nov. 2018, doi: 10.1002/asi.24116.

[28] V. Wang, D. Shepherd, and M. Button, "The barriers to the opening of government data in the UK: A view from the bottom," *Inf. Polity*, vol. 24, no. 1, pp. 59–74, Mar. 2019, doi: 10.3233/ip-180107.

[29] V. Romaniello, P. Renna, and V. Cinque, "A continuous improvement and monitoring performance system: Monitor-analysis-action—Review (MAAR) charts," *IBIMA Bus. Rev.*, vol. 2011, Mar. 2011, Art. no. 917557, doi: 10.5171/2011.917557.

[30] SDAIA. *Open Data Quality Standards Guideline*. Accessed: Jul. 5, 2023. [Online]. Available: https://od.data.gov.sa/downloads/booklet/guidelines/Open%20Data%20Quality%20Guideline_En.pdf

[31] M. Al-Sadani, "Open government data in the Arab world: A survey study proposing a systematic vision," *Arab Fed. Libraries Inf.*, vol. 15, pp. 37–82, Jan. 2015.

[32] P. Iamamphai, J. Noymanee, W. San-Um, and K. Pasupa, "Investigations and comparisons of government open data websites through systematic functional analysis and efficient promotion approach," in *Proc. Manag. Innov. Technol. Int. Conf. (MITicon)*, Oct. 2016, pp. 142–147.

[33] S. Saxena, "Open government data (OGD) in six middle east countries: An evaluation of the national open data portals," *Digit. Policy, Regulation Governance*, vol. 20, no. 4, pp. 310–322, Jun. 2018, doi: 10.1108/dprg-10-2017-0055.

[34] P.-Y. Chu and H.-L. Tseng, "A theoretical framework for evaluating government open data platform," in *Proc. Int. Conf. Electron. Governance Open Soc., Challenges Eurasia*, Nov. 2016, pp. 135–142.

[35] S. Saxena, "Proposing a total quality management (TQM) model for open government data (OGD) initiatives: Implications for India," *Foresight*, vol. 21, no. 3, pp. 321–331, May 2019, doi: 10.1108/fs-07-2018-0073.

[36] V. Wang and D. Shepherd, "Exploring the extent of openness of open government data—A critique of open government datasets in the UK," *Government Inf. Quart.*, vol. 37, no. 1, Jan. 2020, Art. no. 101405, doi: 10.1016/j.giq.2019.101405.

[37] A. Vetro, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to open government data," *Government Inf. Quart.*, vol. 33, no. 2, pp. 325–337, Apr. 2016, doi: 10.1016/j.giq.2016.02.001.

[38] R. A. Sánchez, A. B. Iraola, G. E. Unanue, and P. Carlin, "TAQIH, a tool for tabular data quality assessment and improvement in the context of health data," *Comput. Methods Programs Biomed.*, vol. 181, Nov. 2019, Art. no. 104824, doi: 10.1016/j.cmpb.2018.12.029.

[39] M. Yi, "Exploring the quality of government open data: Comparison study of the UK, the USA and Korea," *Electron. Library*, vol. 37, no. 1, pp. 35–48, Jan. 2019, doi: 10.1108/EL-06-2018-0124.

[40] A. Nikiforova, "Open data quality evaluation: A comparative analysis of open data in Latvia," *Baltic J. Modern Comput.*, vol. 6, no. 4, pp. 363–386, 2018, doi: 10.22364/bjmc.2018.6.4.04.

[41] B. Fan and Y. Zhao, "The moderating effect of external pressure on the relationship between internal organizational factors and the quality of open government data," *Government Inf. Quart.*, vol. 34, no. 3, pp. 396–405, Sep. 2017, doi: 10.1016/j.giq.2017.08.006.

[42] B. Slibar, D. Oreski, and B. Klicek, "Aspects of open data and illustrative quality metrics: Literature review," in *Proc. 35th Int. Sci. Conf. Econ. Social Develop. Sustain. Econ. Social Perspective*, Lisbon, Portugal, 2018, pp. 90–99.

[43] T. Berners-Lee. (2023). *5 Star OPEN DATA*. Accessed: May 7, 2023. [Online]. Available: https://5stardata.info/en/

[44] C. Habernig. *INFOS Cooperation OGD Austria*. Accessed: Feb. 19, 2020. [Online]. Available: https://www.data.gv.at/wp-content/uploads/2013/08/OGD-Metadaten_2-3_2014_11_10_EN.pdf

[45] Ministry of Interior—General Directorate of Traffic. (Jul. 16, 2019). *Traffic Accident Statistics as of 1439 H*. Accessed: Dec. 25, 2019. [Online]. Available: https://od.data.gov.sa/Data/en/dataset/traffic-accident-statistics-as-of-1439-h

[46] I. Al-Turaiki, M. Aloumi, N. Aloumi, and K. Alghamdi, "Modeling traffic accidents in Saudi Arabia using classification techniques," in *Proc. 4th Saudi Int. Conf. Inf. Technol. (KACSTIT)*, Nov. 2016, pp. 1–5.

[47] Yasser. (2019). *Data Quality Guiedlines*. Accessed: Mar. 5, 2020. [Online]. Available: https://od.data.gov.sa/sites/default/files/odp/Open%20Data%20Quality%20Guideline%20V1.0.pdf

NADA FAISAL ALOGAIEL is currently an Instructor with the Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University.

OMER ABDULAZIZ ALRWAIS is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, King Saud University.