جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

“OpenData Insights”

IT 497: Graduation Project Report Product Release-2

Prepared by

Student Name:	Student ID:
Muneera Al-arifi	442200116
Raya Alsuhaim	442201815
Bayan Albadri	442201841
Shadin Alsaif	442200395

Supervised by
Dr. Luluh Aldhubayi

Second Semester 1445
Spring 2024

Table of Contents

List Of Figures

List Of Tables

OpenData Insights

Dr. Luluh Aldhubayi[1], Muneera Al-arifi[2], Raya Alsuhaim[3], Bayan Albadri[4], and Shadin Alsaif[5]

Abstract (English):

OpenData Insights is a web-based application developed to address the issue of poor-quality open data and its impact on decision-making. In today's data-driven world, open data plays a crucial role in various industries, but its quality remains a concern. This software system aims to provide a unified approach to measuring open data quality using 6 international standards. By uploading datasets to the platform, users can obtain comprehensive statistics and figures that reflect the quality of their data based on accuracy, completeness, timeliness, consistency, reliability, and comprehensiveness. The development process followed a systematic software development approach, incorporating user research, interface design, and backend testing. The main findings from the evaluation indicate that OpenData Insights successfully computes open data quality and offers recommendations based on the dataset's domain. The platform's impact lies in empowering users to make informed decisions by utilizing trustworthy open data. It facilitates accurate market analysis, informed investment decisions, and operational efficiencies. Furthermore, OpenData Insights is freely accessible, providing a valuable resource for businesses and government agencies. In conclusion, this software system addresses the fragmented open data quality measurement problem, offering a comprehensive solution to improve the reliability and usability of open data.

[1] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; laldubaie@ksu.ed.sa

[2] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; 442200116@student.ksu.ed.sa

[3] [1]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; 442201815@student.ksu.ed.sa

[4] [1]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; 442201841@student.ksu.ed.sa

[5] [1]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; 442200395@student.ksu.ed.sa

Abstract (Arabic):

تم تطوير تطبيق الويب "OpenData Insights" لمعالجة مشكلة قياس جودة البيانات المفتوحة الرديئة وتأثيرها على عمليات صنع القرار. في عالمنا الحديث الذي يعتمد على البيانات، تلعب البيانات المفتوحة دورًا كبيرًا في مختلف الصناعات، ولكن جودتها ما زالت تشكل قلقًا. لذا، يهدف هذا النظام إلى توفير نهج موحد لقياس جودة البيانات المفتوحة باستخدام ٦ معايير متفق عليها دوليًا عن طريق رفع قواعد البيانات على المنصة، ومن ثم يمكن للمستخدمين الحصول على إحصاءات وأرقام شاملة تعكس جودة بياناتهم بناءً على الدقة والاكتمال والتوقيت، والاتساق، والموثوقية، والشمولية. يكمن تأثير المنصة في تمكين المستخدمين من اتخاذ قرارات مدعومة بالبيانات من خلال الاستفادة من بيانات مفتوحة موثوقة وبالتالي تحسين كفاءة الخدمات الحكومية، وإتاحة الفرص لخلق مجالات عمل وفرص اقتصادية جديدة، والحصول على معرفة جديدة من خلال دمج مصادر بيانات متعددة ومعالجة بيانات ذات كم كبير. علاوةً على ذلك، يتمتع "OpenData Insights" بإمكانية الاستخدام المجاني، مما يوفر مصدرًا قيمًا للشركات والجهات الحكومية. ختامًا، يعالج هذا النظام مشكلة صعوبة قياس جودة البيانات المفتوحة، ويقدم حلاً لتحسين موثوقية واستخدام البيانات المفتوحة.

Keywords: open data quality; software system; decision-making; international quality standards; data analysis; data reliability; data usability.

1. Introduction

Nowadays, open data enables entrepreneurs to drive innovation, penetrate markets, and spot and capture opportunities, it has become a factor in the growth of nations in a variety of areas, including social, environmental, educational, health, and the country's economy. Open data also facilitates decision-making and improves communication between the public and private sectors. It is critical as it provides citizens with the raw materials they need to engage with their governments, become part of improving public services, and participate more in government affairs [1].Therefore, the quality of open data has become a major concern, as the number of opendata sources and the reliance on open data for decision-making has increased. Despite the potential benefits of open data, there are numerous issues and challenges associated with its quality, which limits its effectiveness and limits its usability for various users like business companies. Addressing these challenges surrounding the quality of open data is important to maximize its potential benefits and ensure its effective utilization across various domains. By developing “OpenData Insights“, we can measure the quality of open data and unlock its full potential, this web-based application for assessing open data's quality in accordance with a set of standards will not only assist business owners in making informed decisions but also will expedite the process of verifying open data before making it public, especially for large organizations and government agencies this will save time, effort, and improve companies decision-making process. and by maximizing the economic impact resulting from open data in Saudi Arabia It will contributes to realizing the objectives of Vision 2030. As we are working on developing our web application by focusing on enhancing the Kingdom's utilization of open data. To support the significance of open data, SDAIA has recently released a new tool called " نضيء" on November 27, 2023 [2]. This tool aims to promote the Kingdom's role in open data usage and facilitate decision-making processes. Additionally, it aims to enhance the quality of data,which is an area we are striving to improve upon in our web application.

In this document, you will be acquainted with our interactive open data quality measurement dashboard "OpenData Insights". We will delve into the underlying issue that led us to this innovative solution, in this document, we will provide an overview of our product, focusing on its visionary aspect, scope, and objectives. Additionally, we will outline the roadmap that will guide us toward the final offering, incorporating background research, literature review, system requirements, system design, and testing.

1.1 The Problem

Open data has been a breakthrough in the work of many—tech businesses, policymakers, researchers, etc.—however it has its drawbacks. Although useful open data may be accessible to the general population, Open data quality is crucial for the future of open data because open data is used across practically all sectors and a wide range of fields. However, despite the importance of measuring the open data quality, and the distinct contribution of measurement experiment, some work is duplicated, and there is need for a unified tool to measure and monitor the open data quality.

Low or bad Open Data quality can yield to several issues such as misinterpretation, faulty decision-making, wasted resources and efforts to analyze and use open data, and finally increased costs in data cleaning. It also has a wide impact on the economy such as bad open data can lead to wrong market analysis, poor investment decisions, or operational inefficiencies, all of which can have economic impacts.

1.2 The Solution

“OpenData Insights" is a web-based application that develops an interactive platform measuring the open data quality by uploading the open dataset to the platform. Each uploaded dataset is going through a process to compute its quality and provide statistics and figures according to the 6 international standards of data quality (accuracy, completeness, timeliness, consistency, reliability, and comprehensiveness). Also, the user will be able to compare the standard measurement results of his/her different uploaded open datasets that are stored in his/her profile. Also, users will be able to understand the figures and statistics better since we will use transformer model (ChatGPT) that translates statistics into understandable paragraphs through formatting the statistical data in a manner that the model can interpret. For instance, the open dataset domain will be fed to the ChatGPT, and it will write the suitable standards recommendations on measuring the open dataset also the system will provide the user with similar datasets in that domain. This will enhance decision making in business companies and expedite the process of verifying open data before making it public, especially for government agencies.

1.3 Product

1.3.1 Product Vision

For people who want to make informed decisions based on the open data available. The

"OpenData Insights" is a web-based application that measures the quality of open data according to a set of standards. Unlike other open data measurements applications, our product assesses the open data quality using six open data standards and provides recommendations on suitable standards that measure the quality of the uploaded open dataset, all for free.

1.3.3 Main contribution

Like other websites and applications offering similar services, we've observed that many of them charge fees for access. In contrast, our platform will be completely free to use, and the registration process will be straightforward and intuitive. Beyond just assessing the quality of open datasets, our product will also suggest appropriate standards for uploaded datasets based on their domain. Additionally, it will recommend similar datasets to users, helping them make more informed decisions.

1.3.4 Our Approach

Initially, we conducted extensive user research, engaging with a diverse group ranging from novices to tech enthusiasts, to grasp their needs. Following this, we created simple wireframes to design intuitive interfaces, which we then developed using HTML, CSS, and JavaScript. Leveraging datasets from the Saudi Open Portal, we tested our backend systems to ensure accurate assessment of uploaded data quality. Once all user stories were addressed, we conducted thorough testing with actual users, including individuals from the King Saud University data center, to verify seamless functionality. Additionally, we integrated agile methodology into our process to maintain flexibility and responsiveness throughout the project lifecycle.

1.3.5 Objectives

- **Product (customer focus-value):**

"OpenData Insights" seeks to assist users in making well-informed judgments based on evaluated open dataset and determining whether the open dataset may be utilized in the manner that the user desires. It will assist people in making informed decisions based on trustworthy information, preventing them from regretting the use of poor open dataset in their business plans or any other area that could result in wasted time and money.

This web application will give trustworthy figures such as bar charts and pie charts that are simple for the user to understand, and statistics that are easily visible as a percentage - except for reliability and timeliness and comprehensiveness -, where their results will be displayed as a text since they cannot be measured with numbers, that will enable governments, company owners, and the general public to make the best decisions and use open data, so a wide range of individuals and sectors will benefit from it. The user can do the following:

- Assess the quality of user-uploaded open dataset based on the six international standards (accuracy, completeness, timeliness, consistency, reliability, comprehensiveness).
- Interact with intuitive interface and easily understandable dashboard.
- Create a user profile that stores evaluated open dataset results.
- Get a recommendation on the suitable standards for the uploaded open dataset.
- Get a recommendation of similar datasets of the uploaded open datasets.

- **Project (solution focus-plan):**

1. Define the requirements and identify the quality metrics.
2. Use 8 open datasets for development.
3. Use existing formulas to compute the quality of some of the 6 standards.
4. designing the dashboard.
5. Test the system using 20 open datasets.

- **Learning (student focus):**

- Learn more about the standards of measuring the quality of open data and get on deep with them.
- Improve our skills in python.
- learn how to design dashboards.
- Learn Jira and GitHub more deeply and use it more effectively.
- Learn more about open data.
- Learn how to integrate with ChatGPT and use its API.

1.3.6 Scope

"OpenData Insights" is a web-based application for assessing open data's quality in accordance with a set of international standards which are accuracy, completeness, timeliness, consistency, reliability, and comprehensiveness. It offers certain statistics and figures about the open data that may be useful for individuals who want to use this open data in other areas or enterprises and make educated choices. Users will be able to upload open data and evaluate the reliable nature of it using a set of 6 metrics. They will be able to view graphs, statistics and texts that show the open data quality results. There will be a profile for each user to store his/her results of the uploaded open dataset. Also, users will be able to understand the figures and statistics better since we will use transformer model (ChatGPT) that translates statistics that are sent as numbers, also the standards that are chosen for measurement as a text into understandable paragraphs through formatting the statistical data in a manner that the model can interpret. For instance, the open dataset domain as a text will be fed to ChatGPT and it will write the suitable standards recommendations on measuring the open dataset also the system will provide the user with similar datasets in that domain. There will be an optional sharing functionality to share the results with others. However, the user will not be able to upload more than one open dataset at a time, also only English is supported by the web-based application.

1.3.7 Hardware/Software Tools and Cost

Table 1 represents the tools we used to develop the project.

Table 1 Hardware/Software Tools and Cost.

Hardware Tools	
The project did not require Hardware tools.	
Software Tools	
Name and Description	**Cost**
Jira Software A software application for agile project management that will help our team to manage the project and organize the tasks.	Free

GitHub A software development platform that provides developers with a collaborative environment to work on code, track changes, and coordinate their efforts with the team members.	Free
Flask Flask is a web framework, it’s a Python module that lets you develop web applications easily.	Free
Visual Studio Code An integrated development environment (IDE) that can be used to develop our program's features.	Free
ChatGPT API	Free

There are no other costs required for our project.

2. Background

2.1 Open data

Open data refers to publicly available datasets that are freely accessible and can be utilized, reused, and shared by anyone without any constraints related to technical impediments, financial obligations, or legal restrictions. This inclusive concept applies to the data holdings of both government and non-governmental organizations, as they have the option to openly publish their data, aiming to facilitate broad availability and utilization for various purposes.

Government agencies generate and possess a wide range of data in various sectors. These sectors include population and housing, the organized sector and investment, job market statistics, healthcare, and communications and information technology. Each of these sectors contributes to the diverse types of open data that government agencies produce or own.

In the Kingdom of Saudi Arabia, the National Open Data Platform (od.data.gov.sa) is a key initiative in the national open data strategy. It aims to create a public database that enables transparency and promotes effective participation and innovation within the community. It provides access to a variety of file formats, including XLSX, JSON, CSV, and XML, for download or exploration. Its main role is to display published datasets from ministries and government agencies in an open, usable format. This platform enables end users, such as business owners, to generate, manage, and publish their own open datasets. It also enables governmental and non-governmental organizations to do the same. Its purpose is to enhance community participation, encourage creativity, and make it easier to use the data for research, analysis, and product creation. By providing an efficient and highly reliable user experience that aligns with the best local and international standards, the platform aims to contribute to building a digital economy in the Kingdom [3] [4].

2.2 Data quality assessment

The quality of open data directly affects its usability, credibility, and valuable insights. As data-driven decision-making becomes increasingly important in sectors such as business, government, healthcare, and research, evaluating the quality of data becomes essential. High-quality data plays a critical role in making well-informed decisions and effectively solving complex problems. To ensure this, it is important to create methods and tools that can assess and verify the quality of open datasets before they are shared or used. This will give users confidence

in the data they are working with and allow them to make informed decisions based on accurate and reliable information.

Poor data quality can lead to faulty analyses, flawed conclusions, and misleading decisions. Inaccurate or incomplete data may lead to biased results, unreliable predictions, and inefficient resource allocation. Therefore, it is crucial to evaluate the quality of data before utilizing it. This step helps minimize risks and maintain the integrity of derived insights.

Having a strong understanding of the principles of open data and the six standards for assessing open data quality is essential. The six standards include [5]:

- **Accuracy:** It involves assessing the correctness and precision of the dataset. To calculate the accuracy of a dataset, we will use this formula:

  Accuracy = (total cells -number of cells with errors / number of cells) *100
  according to the formulas found in [6]
- **Completeness:** The extent to which the dataset contains all the required information without any significant gaps or missing values. To determine the dataset's completeness, we shall count the number of blank cells in the loaded dataset, and then divide it by the number of all cells in the dataset, we will use this formula: Completeness = (1-number of incomplete cells/ number of cells) *100 according to the standard definition in [6]
- **Timeliness:** How up-to-date the dataset is and whether it reflects the most recent information available. If the uploaded dataset is from "data.ksu.edu.sa", timeliness can be computed by comparing the last update date of the open data uploaded with the present time by web scraping "data.ksu.edu.sa". If not, the user will be asked to provide the publishing date of this uploaded dataset and then compare it with the present time.
- **Consistency:** The data elements do not contain inconsistencies across different parts of the dataset. The system calculates the Coefficient of Variation (CV) for each quantitative column in the uploaded dataset [6]. And to ensure consistency across qualitative columns, we will perform checks on both the language used and the data type employed.
- **Reliability:** The data is obtained from reliable sources [5]. This is indicated when the sources are real, and authentic, and include the "sa" fragment, as it aligns with the established format for official domains of Saudi entities [7]. After research, we discovered two valuable Python libraries that aid in verifying the reliability of the source. Firstly, the requests library [8] simplifies the process of sending HTTP requests and handling responses, enabling us to assess if the site is real or not. Secondly, the tldextract

library [9] accurately extracts the subdomain, domain, and top-level domain from a URL, facilitating our examination of the resources' top-level domain to ensure authenticity.

- **Comprehensiveness:** The dataset covers the relevant aspects and variables required to make it valuable for its analysis. Since our goal to have a system that deals with open datasets in general, it was hard to find a formula that suits all the domains, and the only person that can decide the importance of the data is the user himself, so after discussing and searching, we decided to make the comprehensiveness measured by using ChatGPT through sending the dataset title, the names of columns, and a snapshot of the dataset. And here is a sample of the prompt to be sent to ChatGPT:

  " For dataset about (title), containing the following columns: X,X. With these rows as a short sample: Answer with Yes or No if this detaset is comprehensive and explain why very briefly.

2.3 Web development

Familiarity with web development will be required to easily design and implement the interactive dashboard. It is the process of building, creating, and maintaining websites using various technologies and programming languages. It involves technical design, coding, and deploying web-based solutions that are accessible through internet browsers.

Web development encompasses both the front-end (client-side) which consists of the user interface (UI) and visual appearance of a website. It involves designing and implementing the elements that users interact with directly. And the back end (server-side) that consists of the components that power a website. It includes databases, logic, APIs, servers, and various other elements that work behind the scenes to support the functionality and data processing of the website [10].

2.3.1 Flask

Flask is a popular and lightweight web framework designed to be simple and used for building web applications in Python. It provides a simple yet powerful foundation for developing web applications. One of the main advantages of Flask is its simple approach, allowing developers to have great flexibility and control over their application's architecture. Flask provides a development server and essential features for web development, such as URL routing, request handling, and template rendering [11].

2.3.2 Dashboards

Dashboards design and development is a specific field of web development that focuses on creating informative and visually attractive interfaces for data visualization and analysis. Dashboards provide users with an at-a-glance graphical display of key metrics, performance indicators, and data insights in a concise and interactive format.

To design and develop dashboards, web developers use a variety of front-end technologies such as HTML, CSS, and JavaScript. These technologies enable the creation of responsive designs, visually attractive layouts, and interactive elements that contribute to an enhanced user experience.

2.3.2.1 Apache ECharts

Apache ECharts is an open-source JavaScript visualization library for creating interactive and customizable charts and graphs [12]. It provides a wide range of chart types and supports various data formats, making it suitable for visualizing data in web applications and other digital platforms. ECharts is freely available for anyone to use, modify, and distribute under the Apache License 2.0.

Some key features of Apache ECharts include [12]:

- Wide variety of chart types: ECharts offers a comprehensive set of chart types, including line charts, bar charts, pie charts, scatter plots, radar charts, heat maps, and more. It also supports combination charts and allows for multiple series and axes in a single chart.
- Interactive and responsive: ECharts provides interactive features like data zooming, tooltip display, and data filtering, enabling users to explore and analyze data within the charts. It also supports responsive design, allowing charts to adapt dynamically to different screen sizes and devices.
- provides configuration parameters for styling elements such as colors, fonts, backgrounds, and borders. Additionally, users can define animations, interactions, and event handling to enhance the user experience.
- Data-driven approach: ECharts follows a data-driven approach, where users can bind data to visual elements in the chart. This enables automatic updates and synchronization between the data and the chart, making it easy to create dynamic and real-time visualizations.
- Plugin system and extensions: ECharts has a plugin system that allows

developers to extend its functionality. There are numerous community-contributed extensions available, providing additional chart types, themes, and interactions.

2.4 ChatGPT API

The ChatGPT API is a powerful tool that provides developers with access to the capabilities of the ChatGPT language model developed by OpenAI. It allows developers to integrate the ChatGPT model into their own applications, products, websites, or services, enabling them to leverage the model's natural language processing capabilities and create interactive conversational experiences. Integrating ChatGPT into our system will play a crucial role in recommending standards to users and suggesting similar datasets based on their uploaded data. By leveraging the OpenAI API, we can harness the power of ChatGPT's advanced AI capabilities to enhance user experiences. Through personalized interactions, ChatGPT can provide tailored recommendations that align with specific domains, helping users navigate and adhere to relevant standards. Additionally, by analyzing the uploaded datasets, ChatGPT can offer valuable suggestions on similar datasets that users may find useful for their projects. The API enables developers to send requests to the model with user input and receive model-generated responses. The API allows for dynamic and interactive conversations by maintaining context across multiple turns, enabling back-and-forth exchanges with the model [13]

3 Literature Review

In this section, we provide a comprehensive review of the existing literature on assessing the quality of open datasets. The review aims to determine the requirements for our project and position it in relation to other related efforts. We organize the literature review based on the key dimensions of dataset quality: accuracy, timeliness, completeness, reliability, consistency, and comprehensiveness.

Poor data quality costs businesses $12.9 million yearly on average. In addition to having a negative immediate effect on revenue, bad data over time makes data ecosystems more complex and results in poor decision-making. By 2022, 70% of enterprises, according to Gartner's prediction, will constantly monitor data quality levels using metrics, increasing it by 60% to significantly reduce operational risks and costs [14]. “Data quality is directly linked to the quality of decision making,” says Melody Chien, Senior Director Analyst, Gartner.

Since information is crucial to online companies, accurate and trustworthy data is a crucial resource on which to base your choices. You'll make the right decisions and outsmart your rivals by doing this [15].

Today, a large number of organizations use various tools to assess the quality of their datasets. They recommended some data quality actions that other businesses should implement, such as the use of data quality dashboards that rely on particular data quality metrics, such as accuracy and consistency [16].

One of these tools is the data science and analytics platform Dataiku, which offers a dashboard for data quality. The dashboard provides metrics and visualizations to monitor and assess the data quality of the platform. Users can submit their datasets and choose how the data quality findings are presented to them. What-if analysis can also be used to interactively test various combinations of input values and examine its effects on expected outputs [17]. Dataiku, however, has a large range of features and functionalities, which can make it rather confusing, especially for users who are new to the platform. To fully utilize its powers, it could need some education and training.

A different tool is Ataccama ONE, which has a data quality feature that evaluates the quality datasets by allowing you to specify data quality rules and applying those rules to your datasets. These rules outline the requirements that your data must fulfill in order to be considered of high quality. To verify data completeness, consistency, accuracy, uniqueness, and adherence to particular data formats or patterns, for instance, you may create rules. A dashboard is used to display and illustrate the findings of the data quality evaluation. You can quickly assess the quality of your data using the dashboard's charts, graphs, and summary statistics. You can create unique data quality criteria in Ataccama ONE depending on your unique requirements. Ataccama ONE, however, is a commercial product, therefore pricing may differ [18].

Another platform that can be mentioned is Talend, it is a useful application that helps evaluate the quality of open datasets. It offers features like data integration, quality assessment, and ensuring data integrity. We looked into Talend's data quality feature and found that it provides summary statistics and graphs on a dashboard to help you identify data quality issues, discover hidden patterns, and spot anomalies quickly. If you need assistance, there are data analysis experts available to help you clean the data. You can also set your own quality standards and receive alerts if the data values exceed certain thresholds. Talend is used by different types of users such as business users, data stewards, data scientists, and data engineers. The process involves submitting the dataset, filling out a form to request data quality evaluation, and then viewing the results on a dashboard. However, because Talend has many features, it can be a bit complex for new or non-technical users [19].

Talend offers both a free version called Talend Open Studio, which is accessible to the general public, and premium alternatives like Talend Cloud and Talend Data Management Platform. The premium versions provide more functionality, scalability, and support options [20].

The platform Informatica Data Quality additionally offers data quality features, allowing for comprehensive analysis of the monitored data. Using the provided data, IDQ's unique visualizations, dashboards, and reports can assist in gaining useful insights. With a drag-and-drop interface, it is user-friendly. It examines the data's completeness, conformity, consistency, duplicates, accuracy, and integrity. Users may, however, add additional user- defined measures that are relevant to the business sector. The most important metrics for the provided dataset and the company can then be specified by IDQ. The results are then shown in a dashboard with figures, and a report is given with a thorough explanation of what's going on [24].

Since informatica Data Quality is a commercial tool that requires proper licensing and

authentication to access its features and functionalities. Users typically need to create an account or obtain a license from Informatica.

Another platform worth mentioning is Alteryx. It offers a wide array of functions, particularly for preprocessing datasets in various formats and extensions. While it does include analytics capabilities, they are not explicitly centered around specific standards like our platform, where users are constantly reminded that their datasets are being assessed based on six international standards. Alteryx does provide tools for delving into trends and patterns using low-code, no-code spatial, and prescriptive analytics features, but these may require users to be familiar with technical terms. Similar to the others mentioned, Alteryx is a paid platform, necessitating users to pay fees for its usage.

Upon reviewing similar platforms as represented in the following table [Table 2], we observed a common trend where most web applications require payment and involve complex registration processes before users can access their features. Despite offering extensive functionalities tailored for large enterprises, these platforms present accessibility challenges. In response, "OpenData Insights" aims to simplify access by providing a cost-free solution for all users. Our approach involves evaluating data quality against six internationally recognized standards, presenting transparent statistical analyses, seeking feedback for continuous improvement, and delivering comprehensive summary reports. Our interface will prominently display easily understandable numerical insights, akin to competitors such as Informatica Data Quality. Furthermore, we will offer standardized recommendations based on dataset characteristics. In contrast to competing platforms, our registration process will be swift and without financial obligations, ensuring accessibility for all users.

Table 2 Competitors Summary.

Platform /Feature	Summary statistics and graphs	Reports Generation	Free version available	Straightforward registration
Dataiku	No	No	Yes	No
Ataccama ONE	Yes	No	No	No
Talend	Yes	Yes	Yes	No
Informatica Data Quality	Yes	Yes	No	No

Alteryx	No	No	No	No
OpenData Insights	Yes	Yes	Yes	Yes

Additionally, we discovered that it is tough and complex to calculate reliability, timeliness, and comprehensiveness; nonetheless, after evaluating and searching through published datasets, we discovered that there are several trustworthy and reliable sites, including:

- data.gov.sa
- data.ksu.edu.sa

We developed our own approach for determining comprehensiveness and timeliness after merging the most crucial elements and taking them into account because there are several ways for assessing them depending on the dataset type and domain.

According to the literature, there is no defined threshold on data quality. To say that data is of a good quality, it must have 100% in every dimension of quality standards. For example, if your data contains 20/100 cells inaccurate data, then the data accuracy is only 80%, and therefore 20% of the data is not in a good quality. We can say that, in total, any score below 100% indicates data is not good quality data.

3.1 Competitive Product Analysis

Talend is a useful application that helps evaluate the quality of open datasets. It offers features like data integration, quality assessment, and ensuring data integrity. We looked into Talend's data quality feature and found that it provides summary statistics and graphs on a dashboard to help you identify data quality issues, discover hidden patterns, and spot anomalies quickly. If you need assistance, there are data analysis experts available to help you clean the data. You can also set your own quality standards and receive alerts if the data values exceed certain thresholds. Talend is used by different types of users such as business users, data stewards, data scientists, and data engineers. The process involves submitting the dataset, filling out a form to request data quality evaluation, and then viewing the results on a dashboard. However, because Talend has many features, it can be a bit complex for new or non-technical users [19].

Talend offers both a free version called Talend Open Studio, which is accessible to the general public, and premium alternatives like Talend Cloud and Talend Data Management Platform. The premium versions provide more functionality, scalability, and support options [20].

Dataiku is another platform that allows users to upload data of any format in any size. Users can ask inquiries using the self-service feature and interactive dashboard without having to wait for an experienced staff to respond. This platform also offers "what-if" scenarios, which boosts confidence between data providers and business stakeholders [21]. Additionally, Dataiku gives users the option to choose whether they want to do data quality checks on the entire datasets or just selected samples for better results [22]. This data quality assessment can be based on a variety of variables, including statistical summaries that highlight outliers and missing values, user-defined criteria and requirements that the dataset must meet, or even data quality metrics like completeness, accuracy, and consistency. Although the platform has a user- friendly layout, non-technical users would need some time to fully understand all of its features [23].

Dataiku offers both of the platform's editions: Dataiku Enterprise Edition, which costs money, and Dataiku Community Edition, which is available for free. The Enterprise Edition offers features and capabilities that are appropriate for bigger teams and organizations while the Community Edition has some restrictions on data storage and distribution.

Both Talend and Dataiku allow users to register for accounts; however, with Talend, registration is often only necessary for accessing Talend Cloud rather than the software directly. In order to access and use the data quality features on Dataiku, users normally need to register for an account or profile on the website.

The platform Informatica Data Quality additionally offers data quality features, allowing for comprehensive analysis of the monitored data. Using the provided data, IDQ's unique visualizations, dashboards, and reports can assist in gaining useful insights. With a drag- and-drop interface, it is user-friendly. It examines the data's completeness, conformity, consistency, duplicates, accuracy, and integrity. Users may, however, add additional user- defined measures that are relevant to the business sector. The most important metrics for the provided dataset and the company can then be specified by IDQ. The results are then shown in a dashboard with figures, and a report is given with a thorough explanation of what's going on [24]. Since informatica Data Quality is a commercial tool that requires proper licensing and authentication to access its features and functionalities. Users typically need to create an account or obtain a license from Informatica.

After looking for similar competitors, we discovered several shared traits, such as the fact

that the accuracy, completeness, consistency, and reliability of international standards are typically used to evaluate the dataset quality. Additionally, the outcomes are frequently shown on a dashboard with numbers and statistics. Some competitors, including Informatica Data Quality, moreover, offer a report that details the results of the analysis. One of the features we'll offer is comparable to this one. Depending on the dataset provided and its contents, "OpenData Insights" will also provide standard recommendations. We will make our application available to all of its users for free with extremely straightforward procedures when making an account because the majority of the similar tools we found require some lengthy registration process and may require some money to utilize it.

Table 3 Competitive Product Analysis.

Competitive Product Analysis			
	Talend	**Informatica (IDQ)**	**Dataiku**
Supports Open datasets analysis?	Yes	Yes	Yes
Paid or free?	Both	Paid	Both
Supports Arabic datasets analysis?	Yes	Yes	Yes
Can a user create a profile?	Yes	Yes	Yes

4 System Design and Development

4.1 Methodology

4.1.1 Agile Approach

To develop 'OpenData Insights' website we have used agile approach, which enabled us to deliver incremental updates and foster an iterative and responsive development process. This approach ensured the website's continuous evolution based on valuable user feedback and dynamic requirements. By conducting continuous testing throughout the development process, we identified and addressed issues at their earliest stages.

The incremental development process promoted frequent communication between team members, with the goal of achieving a common understanding of project goals and maintaining continuous progress. Our collective efforts in this environment have been dedicated to delivering

a website that is robust, deployable, and aligned with our stakeholders' expectations.

4.1.2 Scrum Framework

Scrum is widely recognized as the most popular and commonly used agile framework. It enables teams to adapt quickly to changes and helps them to deliver the work quickly and periodically.

Scrum has a team which is composed of three key roles: Scrum master, Product owner, and Developers, each with specific accountabilities and responsibilities aimed at turning work into valuable increments during a Sprint.

The Scrum Master is responsible for helping the team to build the product by guiding them to use scrum, embodying agile principles, and ensuring that the team follow these guides. The Product Owner is tasked with formulating and communicating the product's goal, prioritizing the work that needs to be done, and ensuring to deliver results that align with user's needs. The Developers are responsible for the actual development and delivery of the product, they are the ones who build the product and deliver the work [25]. Our scrum team is shown in Table 4.

Table 4 Scrum Team.

Scrum Team	
Product Owner:	Dr. Luluh Aldhubayi
Developers:	Raya Alsuhaim Bayan Albadri Muneera Al-arifi Shadin Alsaif
Scrum Master (SM):	Dr. Luluh Aldhubayi
Stakeholders:	Gp2 committee

The scrum framework consists of 3 roles, 5 events, and 3 artifacts. Previously we have mentioned what are the roles and their responsibilities. The 5 events are: Sprint, Sprint Planning, Daily Scrum, Sprint Review, and Sprint Retrospectives. Throughout the Sprint, the Scrum team works collaboratively to complete the tasks identified in the sprint backlog and deliver valuable outcomes. Sprint Planning occurs at the beginning of each sprint. It involves the team and Product

Owner planning the work to be done in the upcoming sprint, reviewing, selecting items from the product backlog, and defining the sprint goal. The Daily Scrum is a brief daily meeting -around 20-30 minutes- where we, the developers, discuss the progress, plan the work for the day, share updates to each other, and identify any obstacles we may encounter. At the end of each sprint, we do a Sprint Review, where the team presents the completed work accomplished during the sprint to stakeholders to gather feedback and make any adjustment to the product backlog. The Sprint Retrospective is a meeting held after the Sprint Review where the team examine their processes, identify strengths and areas for improvement, and create a plan for implementing those improvements in the next sprint. The 3 artifacts are: Product Backlog, Sprint Backlog, and Product Increment. The Product Backlog serves as a prioritized list of features and requirements necessary to complete the project. It provides a comprehensive catalog that guides the development process. The Sprint Backlog, on the other hand, consists of the specific items selected from the Product Backlog for completion during the current sprint. It represents the work that the team commits to delivering within the sprint. Finally, the Increment encompasses all the completed work from the current sprint, as well as all the previous increments. It represents the tangible outcome of the team's work [26].

With adherence to the Scrum framework and practicing Agile principles, and by following scrum's roles, events, and artifacts. Our team aimed to deliver value incrementally, involve stakeholders throughout the development process, and continuous feedback gathering to ensure that the product aligned with the user's requirements. Scrum provided our team with flexibility and adaptability as new insights emerged, enabling our team to respond to changes effectively.

4.1.3 Tools

During the implementation of our system, we utilized both GitHub and Jira. These tools played a big role in our incremental development process, which involved continuous monitoring and updating. GitHub[6] facilitated seamless collaboration on our codebase and enabled us to track changes made by team members. Its functionality allowed us to work concurrently on the same code without conflicts, while its reviewing and merging features simplified the code integration process. Furthermore, Jira[7] served as an effective task management tool, enabling us to prioritize and assign tasks among team members. It also provided a means to monitor our progress throughout the project.

[6] GitHub: OpenDataInsight/2023-GP1-7-Final-Release-1 (github.com)
[7] Jira: PNDT board - Agile board - Jira (atlassian.net)

4.2 System Requirements

4.2.1 System Users

“OpenData Insights" has 2 types of users:

1- **Data publishers:** Individuals over 20 years old or organizations who publish open datasets and want to ensure their datasets meet certain quality standards. They have a high level of technical expertise, particularly related to data analysis, data quality, and data management. They are comfortable working with data manipulation tools, programming languages, statistical software, and data visualization tools. They have a good understanding of data structures, data formats, and data processing techniques.

2- **Data consumers:** Users who rely on open datasets for their research, analysis, or decision-making process and need to verify the dataset's quality. They typically have a higher educational level, including individuals with undergraduate and graduate degrees. They may have backgrounds in various fields such as computer science, statistics, social sciences, engineering, or other related disciplines. They may have varying levels of experience in working with open datasets. Some might be novice or entry-level professionals who are just starting their careers, while others may have years of experience in their respective fields.

4.2.2 Requirements Elicitation

During the elicitation of requirements for our project, the following sources of information were utilized:

1. **Documentation:** Relevant documents, such as previous research papers, industry reports, and best practices guides, were studied to gather insights into the challenges and requirements related to open dataset quality assessment.

   Stakeholders: Input and feedback were gathered from stakeholders involved in open data initiatives, including experts and data consumers. Their knowledge and expertise were instrumental in understanding the current practices, challenges, and expectations regarding dataset quality assessment.

The following requirement discovery methods were employed to gather information and insights:

1. **Interviews:** Interviews were conducted with experts and data consumers to understand their experiences, challenges, and preferences related to open dataset quality assessment. These interviews provided valuable qualitative data and allowed for in- depth discussions.

2. **Questionnaires:** Questionnaires were distributed to experts and data consumers to gather data and obtain a broader perspective on their practices, expectations, and opinions regarding dataset quality assessment. The questionnaires provided a structured approach to collect information from a larger sample size and were posted online using Google forms and collected data for a week starting on September 11, 2023.

3. **Focus Groups:** Focus group discussions were conducted with relevant stakeholders to facilitate interactive discussions and gather diverse perspectives on open dataset quality assessment. These sessions allowed for group dynamics and collaboration among participants.

4. **Expert Consultation meetings:** We sought advice and input from domain experts and consultants such as the Data Governance Office at SDAIA and the Data Governance Office at King Saud University who have extensive knowledge and experience in the field of open data. Their experience provided valuable guidance that helped define some important requirements.

   During our interview with SDAIA, we discovered that the data quality assessment process is primarily manual and limited in scope. However, SDAIA has made progress in this area by developing a project that focuses on measuring data quality. It's important to note that this project is currently applicable to datasets in general and primarily available to agencies responsible for publishing data, such as governments. SDAIA believes that the responsibility for checking data quality lies with the agencies themselves, as the existing process is basic, as previously mentioned. When discussing standards, SDAIA emphasized that assessing the comprehensiveness of datasets should be based on their specific domains. In terms of accuracy, they informed us that there is no standardized process or method for measuring it. Additionally, the concept of thresholds was discussed, and they agreed that general thresholds applicable

to all datasets are not readily available.

The office of data management at King Saud University operates solely through manual processes, without utilizing any tools. One notable challenge is the high cost and annual license renewal requirements associated with popular data quality measurement tools. However, Professor Ibrahim advocates for the development of a local tool, expressing optimism about its potential widespread acceptance among government entities. The determination of data quality thresholds lies with the data requester, who specifies the relevant fields. Among these standards, measuring timeliness proves to be particularly difficult, as a universally accepted method has not yet been established. To ensure the quality of open data, it is crucial to provide descriptive metadata, which greatly facilitates the evaluation process.

Through these sources and requirement discovery methods, a comprehensive understanding of the challenges, expectations, and preferences regarding open dataset quality assessment was obtained. This information served as the foundation for defining the requirements and designing the features of the quality assessment tool.

The details of this questionnaire are listed in **APPENDIX A:** Expert's questionnaire.

As a result of the experts' questionnaire, the following results have been obtained:

- The majority of experts (60%) have higher education, while 38% have bachelor's degree.
- About 88% of experts take a lot of time measuring dataset quality.
- About 52% of experts rely on open datasets for their work on a(Daily/Weekly/Monthly) basis, while 44% on a Rarely Basis.
- The majority of experts (60%) prefer both statistics and figures while 32% prefer figures alone.
- About 56% of experts are likely to recommend high-quality open dataset results with others while 40% are Neutral.
- About 56% of experts are likely to switch to alternative open datasets if they

encounter quality issues with their dataset.

- The majority of experts (96%) rely on visualization tools or dashboards to explore and analyze open datasets.

Listed in **APPENDIX B:** Expert's interview.

From their answers, we concluded:

- Even experts often find ensuring the quality of the open datasets challenging and time consuming.
- All experts expect the process to be efficient and take a reasonable time to process the dataset and be presented with results/recommendations in not more than 3-5 minutes.
- All the experts interviewed would love to have a final quality score for each Standard.

Listed in **APPENDIX B:** Expert Consultation meeting.

During our meeting with The Saudi Authority for Data and AI (SDAIA), we discovered that the data quality assessment process is primarily manual and limited in scope. However, SDAIA has made progress in this area by developing a project that focuses on measuring data quality. It's important to note that this project is currently applicable to datasets in general and primarily available to agencies responsible for publishing data, such as governments. SDAIA believes that the responsibility for checking data quality lies with the agencies themselves, as the existing process is basic, as previously mentioned.

When discussing standards, The Saudi Authority for Data and AI (SDAIA) emphasized that assessing the comprehensiveness of datasets should be based on their specific domains. In terms of accuracy, they informed us that there is no standardized process or method for measuring it. Additionally, the concept of thresholds was discussed, and they agreed that general thresholds applicable to all datasets are not readily available and can vary based on the domain of the open dataset.

And for our meeting with the office of data management at King Saud University, the office operates solely through manual processes, without utilizing any tools. One notable challenge is the high cost and annual license renewal requirements associated with popular data quality

measurement tools. However, Professor Ibrahim advocates for the development of a local tool, expressing optimism about its potential widespread acceptance among government entities. The determination of data quality thresholds lies with the data requester, who specifies the relevant fields. Among these standards, measuring timeliness proves to be particularly difficult, as a universally accepted method has not yet been established. To ensure the quality of open data, it is crucial to provide descriptive metadata, which greatly facilitates the evaluation process.

Listed in **APPENDIX C:** Data consumer's questionnaire.

As a result of the Data consumer's questionnaire, the following results have been obtained:

- The majority of Data consumers (60.6%) have higher education, while 34.2% have High School Diploma or equivalent.
- About 60.6% of Data consumers rely on open datasets for their work on a(Daily/Weekly/Monthly) basis, while 36.8% on a Rarely Basis.
- About 65.8% of Data consumers are likely to recommend high-quality open dataset results with others while 28.9% are Neutral.
- About 57.9% of Data consumers find it difficult to ensure that open dataset is reliable and of good quality while 34.2% said they "sometimes" find it difficult.
- About 55.3% of Data consumers are likely to switch to alternative open datasets if they encounter quality issues with their dataset while 39.5% are Neutral.
- 97.4% of Data consumers would like for the system to recommend similar datasets to the one they upload.
- The majority of Data consumers (86.4%) prefer both statistics and figures while 26.3% prefer figures alone.
- 68.4% of Data consumers said they would be interested in different ways (like a brief description) to help them understand the statistics and figures while 31.6% said maybe.

Listed in **Appendix E:** Data consumer's interview.

From their answers, we concluded:

- Some users would love to compare multiple datasets through visual representations, allowing users to analyze and evaluate the datasets effectively.
- a history feature would be beneficial, allowing users to maintain records of previous assessments for future reference.
- All the Data consumers interviewed prefer figures explaining information.
- Some Data consumers interviewed rely heavily on datasets approved by SDAIA (Saudi Data and Artificial Intelligence Authority) but encounter limitations in finding the specific datasets they require.
- Ensuring the reliability and quality of open datasets before utilizing them for decision-making or research purposes is a significant challenge, as expressed by some Data consumers interviewed. This difficulty can have a direct impact on the user's work and productivity.

4.2.3 User Interactions

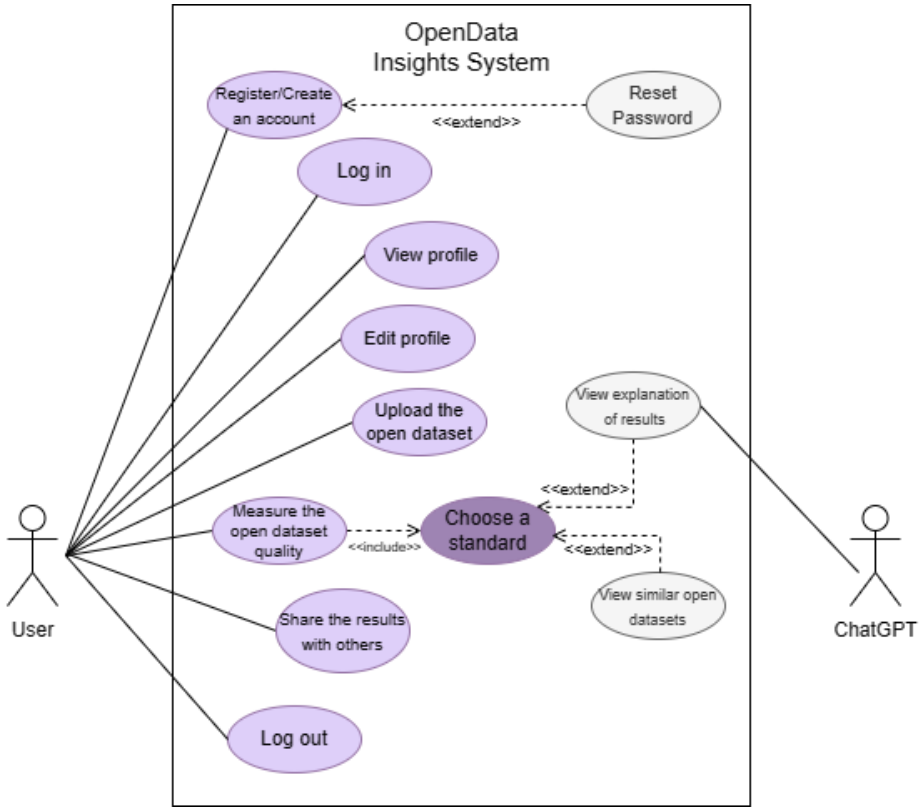


Figure 1 Use case diagram.

4.2.4 Product Backlog and Roadmap

4.2.4.1 Product Backlog Table

In this product backlog shown in Table 5, we have listed all the functionalities of our product based on their priority.

Table 5 Product Backlog.

ID	PBIs (User Stories)	Size	Type	Status	Acceptance Criteria
1	As a user, I want to register so that I can see all the website services and interact with the system.	3	Feature	Done.	The registration page must include a form to enter information and a submit button (Sign Up Now), after clicking this button the system must check that the user does not have a previous account based on his/her email. If it does have an account a message will appear, saying "Email is already in use. Please log in instead". If it does not have an account, the user will be directed to the system home page. The system should check if the entered email is in an email format or not. If not, a message will appear, saying "Invalid email format".

The system should check if the entered password is 6 characters long or not. If not, a message will appear, saying “Password should be at least 6 characters long”.

- The system should check if the entered password and the confirmed password match or not. If not, a message will appear, saying “Password and Confirm Password do not match”.

- The system should securely store the user's login credentials (email and password) after successful registration.

- Registered users should have access to all the available system features and functionalities.
- User credentials and sensitive information should be protected by password hashing and encryption.

- Registration should be straightforward, intuitive, and easy to understand. The user interface should provide clear instructions and guidance to assist users in completing the registration successfully.

2	As a user, I want to log in so that I can access my account.	5	Feature	*Done.*	- When the user enters their account information in the login form and clicks the "Login" button, the system should verify if their account exists. - If the account exists, the system should allow the user to log in and direct them to the home page. - If the account does not exist or he/she entered wrong credentials, then an error message is displayed.
3	As a user, I want to be able to reset my password so that I can access my account even if I forget my old password.	3	Feature	*Done.*	- When the user clicks “forgot password” button from login page, he should be directed to “reset password” page. - A form should be displayed; the form should contain a single input field for entering the email address associated with the user's account. - When a user fills in their email address and submits the form by clicking the "Send Reset Email" button, the form should trigger a password reset email to be sent to the provided email address. - If the email field is left empty upon submission, an error message should be displayed indicating that the email field is required. - If an invalid email address format is entered (e.g., missing "@" symbol), an error message should be displayed indicating that the email address is invalid. - If the provided email address does not match any existing

					account in the system, an error message should be displayed indicating that no account associated with that email address was found. - Upon successfully sending the password reset email, a success message should be displayed informing the user that the password reset email has been sent successfully.
4	As a user, I want to log out so that I can exit my account.	2	Feature	*Done*	The system should provide a clearly visible “Log out” button. - When the user clicks on the "Log out" button, the system should log them out and redirect them to the login page.

5	As a user, I want to be able to upload the open dataset so that I can view its quality assessment results.	3	Feature	Done.	- When the user clicks on the upload area, he/she should be able to choose an open dataset from his/her device. - The system should accept the following extensions: (XLSX, JSON, CSV, and XML) After clicking on the "Upload" button: - The system should provide a message indicating that the open dataset has been uploaded successfully. - If the user provides an open dataset with a non-supported extension, then a message will appear

					stating that the uploaded open dataset format is not supported. - The dataset must be saved in a web directory. - It should be easy to understand how to upload, and the upload button should be clearly visible.
6	As a user, I want the system to calculate the completeness standard of my open dataset so that I can see the result of the quality assessment according to this standard.	5	Feature	Done.	- First the system should accurately calculate the completeness through the following formula: [6] Completeness = (1-number of incomplete cells/ number of cells)*100 - To determine the dataset's completeness, we shall count the number of blank cells in the loaded dataset. - If the dataset has not been assessed for completeness yet, the

					completeness standard section should indicate that the assessment is pending. - If the dataset has been assessed for completeness, then the percentage of the complete data cells should be displayed on the dashboard as an interactive figure and static(percentage). The percentage of the incomplete data cells should be shown in the interactive figure.
7	As a user, I want to view my profile information so that I can compare between my uploaded open datasets quality results	2	Feature	Done.	- As a user, If I click on "profile" icon, the system should direct me to profile page, then I should be able to view my information such as my email, my name, and my previous uploaded open dataset's name, and domain. - When clicking on one of my open datasets, it should direct me to the

					dashboard and display the result of the open dataset. However, if the user is a new user and has not uploaded any datasets yet, only the email and name should be displayed.
8	As a user, I want to edit my profile information such as my password so that I keep my account secure.	3	Feature	Done.	- When I navigate to the profile page, I should be able to change password and to delete an uploaded open dataset. - Once I click on the "change password" button next to the password field, a form should appear that includes the old password field, new password field for entering the new password, and Confirm Password field to confirm the new password. - If I leave the password field blank, I should receive an error message indicating that the password field is required. - If I entered a wrong old password a "wrong password"

					message should appear. - If I enter a new password that does not meet the specified password requirement (6 minimum length), I should receive “Password must be at least 6 characters!” error message. - If I enter valid value for the password field and fill in the form then click the "Save" button, my password should be updated, and I should receive a success message indicating that my password has been changed. - I should see a button or icon indicating the delete on the uploaded open dataset. - If I click on the delete icon/ button a conformation message should appear to confirm the deletion.

					If I agree on the deletion process the uploaded open dataset should be deleted.
9	As a user I want to choose the standard measure that I want, so that I can view the quality assessment result according to the chosen standard.	5	Feature	Done.	- A checkbox will appear with the standards as the choices labeled as "Reliability" "Consistency" "Completeness" "Comprehensiveness" "Timeliness" and "Accuracy". So, the user can choose which standard he/she wants to measure. - A “Start” button will be disabled until the user choses at least one standard. - Upon selecting a specific standard measure, the system updates the quality assessment result accordingly.
10	As a user, I want the system to recommend me which standards to choose, so that I can get the best quality assessment results.	5	Feature	Done.	- The user will be asked to choose a domain from a drop-down menu.

					- The system will send the domain chosen to ChatGPT, different domains may have specific standard measures or assessment criteria that are more relevant or commonly used within that domain. - The chosen domain will be sent as a text prompt to ChatGPT in order to recommend suitable standards for this uploaded dataset. then, the system will receive the recommended standards for the dataset. - The recommended standard measures should be clearly presented to the user as an easily understandable message.
11	As a user, I want the system to tell me about the reliability standard of my open dataset so that I can see if I should trust the origin of this uploaded dataset.	5	Feature	Done.	- The system should ask the user to provide the links of the source of the open dataset (from

					where he/she got it), and the author of the open dataset. - The system should check the provided links for two conditions, first if they are valid and accessible, by performing an HTTP request and examining the status code. The link is valid if the status code falls within the range of 200 to 399, inclusive [8]. Second the system should check the provided links, if the top-level-domain of the links is "sa" by using the tldextract library [9]. - If the links fulfil the two conditions, it's considered reliable, and the result will be the word "Reliable" in green color. - If the links don't fulfill the two conditions, or one of them, then the result will be the word "Unknown" in grey color.

12	As a user, I want the system to calculate the timeliness standard of my open dataset so that I can see the result of the quality assessment according to this standard.	5	Feature	Done.	- When the timeliness standard is chosen: 1- An input field will be displayed to the user asking them to provide the website link they got the open dataset from. 2- Another input field will be displayed to the user asking them to provide the number of years and months that will be used in the comparison function. - If the uploaded dataset is from data.ksu.edu.sa, then timeliness can be computed, and the system will compare the current date and the publishing date of the open dataset uploaded by web scraping data.ksu.edu.sa. - If the provided dataset is not from data.ksu.edu.sa, the

					user will be asked to provide the publishing date of this uploaded dataset. Then the system will compare the current date with the publishing date. - If the difference of years and months is x years and y months or more - where x and y are the provided integers input from the user-, then the system will display “out of date dataset" to the user. Else “Up to date" is displayed.

13	As a user, I want the system to calculate the consistency standard of my open dataset so that I can see the result of the quality assessment according to this standard.	5	Feature	Done.	- When consistency is chosen, the system should calculate the Coefficient of Variation (CV) for the uploaded dataset quantitative columns. - The CV should be calculated using the formula: CV = (Standard Deviation / Mean) * 100 - The CV calculation should be performed for relevant numerical or quantitative attributes within the dataset. - For qualitative attributes the system should check the consistency of the language used to ensure that it's only one language used per column. - For qualitative attributes the system should check the consistency of the data type to ensure that it's only one data type used per column.

					- For the quantitative columns, the results should be displayed as a percentage % with bar charts that visually represent the percentage results. - For qualitative columns the system should show a table that has the result as a text, “Yes” when the column is consistent, and “No” when the column is not consistent.
14	As a user, I want the system to calculate the accuracy standard of my open dataset so that I can see the result of the quality assessment according to this standard.	5	Feature	Done.	- When the user chooses accuracy, the system will calculate the accuracy based on pre-defined criteria and file extension. - It converts the file to a data frame as it will consider both qualitative and quantitative columns of the dataset. - If the dataset contains qualitative columns: - The system will check if the column is found in the reference file which contains most common attributes between datasets in the same domain. Discrepancies between the

					dataset values and the reference values will be identified and considered an inaccurate cell. - If the qualitative column is not found in the reference file, then each of its unique values will be considered an inaccurate cell. - If the dataset contains quantitative columns: - The system will calculate mean and standard deviation for each numerical column. Values deviating from the mean by a specified threshold multiple of the standard deviation will be flagged as inaccurate. - The accuracy is finally calculated based on the following formula: Accuracy = (total_cells - incorrect_cells) / total_cells - It will then be displayed as a percentage and a bar chart to the user.
15	As a user I want to see similar datasets to the one I uploaded so that I can discover additional attributes that are present in the similar datasets but not in mine for data improvement.	5	Feature	Done.	- When the user clicks on the" Show suggestion" button the system should suggest three out of five open datasets that are saved in the system randomly. - The open datasets must be from the same domain as the uploaded open

					dataset. - The suggested open datasets should be shown as a URL link, the system should direct the user to the open dataset in open data platform when user clicks on it.
16	As a user I want to see a report that summarizes the results of the figures and statistics so that I can get a better understanding.	5	Feature	Done.	- When the user clicks the "Generate Report" button on the dashboard page, the system should generate a report summarizing the results of the assessment figures and statistics. - A prompt will be sent to ChatGPT with the results of the chosen standards by the user. - If any of the values for accuracy, completeness, reliability, timeliness, comprehensiveness, or consistency is not available (i.e., "None"), it should be mentioned in the report that the information is not available or not mentioned. - A paragraph will be shown back to the user with a description of the result from ChatGPT as an answer to the prompt sent.

17	As a user, I want the system to calculate the comprehensiveness standard of my open dataset so that I can see the result of the quality assessment according to this standard.	5	Feature	Done.	- When the user chooses comprehensiveness, the system should connect to ChatGPT and send a prompt including the open dataset “title”, “the names of columns in the uploaded open dataset” , “and a snapshot of the open dataset”. - The prompt should be as follows: For dataset about (title) containing the following columns:” column names”. With these rows as a short sample: Answer with Yes or No if this dataset is comprehensive and explain why very briefly - The system should present the result as a text. “Yes” or “No” to indicate whether it’s comprehensive or not, then a text to explain why.

18	As a user, I want to be able to share the data quality results so that I can refer to it whenever I want.	5	Feature	Done.	- The system provides a "Share" button. - The "share" button will be disabled unless the user downloads the dashboard by clicking "Download" button. - When the user clicks the "Download" button, a pdf file will be generated including the content of the dashboard and then it will be downloaded in the users' devices. - After downloading the dashboard, the user can click the "Share" button. - After clicking the "Share" button, the system will trigger the users to the default email application on their device allowing them to send emails to whoever they want.

19	As a user, I want the system to be fast in response, within 10 seconds, so that I can quickly assess the quality of the open dataset without experiencing delays. (performance)	5	Feature	Done.	- When the user enters the system and performs any actions the response time must be within 10 seconds. - The system should response to any action in not more than 10 seconds. - The system should response within the same regardless of the specific action they perform.
20	As a user, I want to intuitively understand and use the system within 10 mins without requiring special training so that I don't waste time learning and make any mistakes. (usability)	5	Feature	Done.	- The user should be able to understand the system's main features, layout, and functionality within 10 minutes of exploring it for the first time. - The system should have clear and obvious icons and labels that help the

					user while using the system. - The system should not cause the user to make mistakes in more than 10% of the user interactions.
21	As a user, I want the system to be reliable 95% of the time when the user numbers increase so that I can access and use the system without degradation or downtime. (scalability)	5	Feature	Done.	- The system should scale horizontally. - The system should perform the same performance regardless of the number of users.
22	As a user I want to be able to use the system on multiple devices so that I can access it whenever I want at any device (compatibility)	5	Feature	Done.	- I should be able to access and use the system in any device. - The system should be responsive and adapted to different screen sizes and resolutions. - All system features and elements should be accessible and usable on different devices without any loss of functionality or visual integrity. - The system should be responsive and compatible to these most used web browsers (Google

					Chrome, Safari, Microsoft Edge)
23	As a user, I want open data insights to be available 99% of the time I try to access it, so that I can access and evaluate my open dataset.	5	Feature	Done.	- Open data insights must be available 99% of the time. - The system should be unavailable only 1% of the time, and that's during the system downtime such as during maintenance, unplanned failures or the time it takes for a system to recover. - The availability percentage should be regularly monitored through using "UptimeRobot" tool and measured through this formula Availability = Uptime ÷ (Uptime + downtime) [27]

4.2.4.2 Roadmap

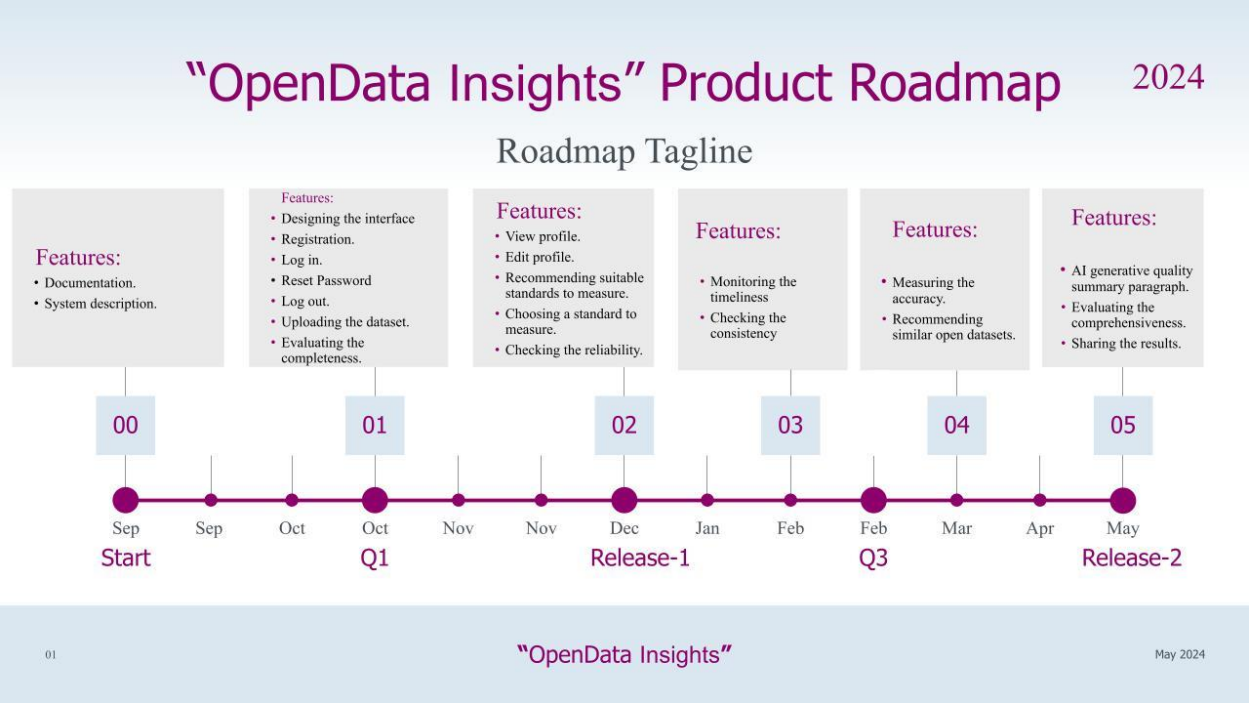


Figure 2 Roadmap.

4.2.4.3 Definition of Ready

✓	The acceptance criteria should clarify the formula used for every standard.
✓	The acceptance criteria should clarify the accepted dataset format like (Excel, JSON, CSV, and XML).
✓	User stories should be discussed with the team and understood by them.
✓	User stories priorities are clear to the team.
✓	User stories should be organized in the backlog based on their priority.
✓	The team should provide an initial estimation of the effort and the size of each user story.
✓	The size of the user stories should be small enough to be completed in a sprint.
✓	The acceptance criteria of a user story should be clear, precise and meet the conditions for a user story to be accepted.
✓	The user stories are clear and precise.

4.3 System Design

4.3.1 Architectural Diagram

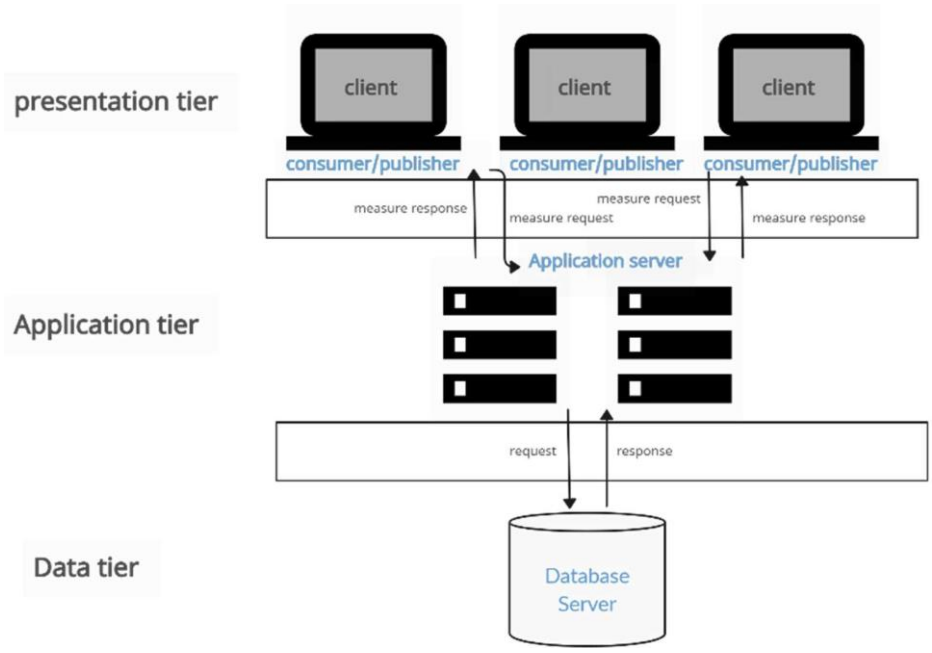


Figure 3 Architectural design.

Our system is a web-based application that enables many users to request and receive services from a centralized server.

The figure shows the main components in the Tier-3 client-server architecture:

- The presentation tier, also known as the client tier, is responsible for handling the user interface and user interaction. It sends requests to the application tier and receives responses to display the data and results to the user.
- The application tier handles tasks such as user authentication, data validation, dataset analysis, standard recommendations, and report generation. It communicates with the data tier to fetch or store data as needed.
- The data tier is responsible for managing data storage and retrieval. It includes databases where datasets, user account information, analysis results, and other relevant data are stored.

Because of its architecture, clients can visit the website through a network, and it can receive and process many requests from numerous clients and servers at once.

The following limitations prevented us from utilizing the other architecture design patterns for our website:

- MVC Architecture: This design pattern is applied when there are numerous ways to view and interact with the data. When the data model and interface are simple, it also requires extra code and data complexity. Because of the simple nature of our website, MVC design is not appropriate because we do not require different views.

- Layered architecture is frequently utilized in circumstances when performance is not a top priority. Performance is a top priority in our system, and layered architecture can make it worse.

- Repository architecture is used when a lot of data needs to be shared and the components may function independently. Because our website's components depend on one another and must communicate with one another, repository architecture is inappropriate for it. Repository architecture is also a single point of failure; thus, issues there affect the entire system.

- Pipe and filter architecture is employed in non-interactive systems, processing the input in steps to produce the output. The pipe and filter design is inappropriate because our website requires an interactive system

4.3.2 Class Diagram

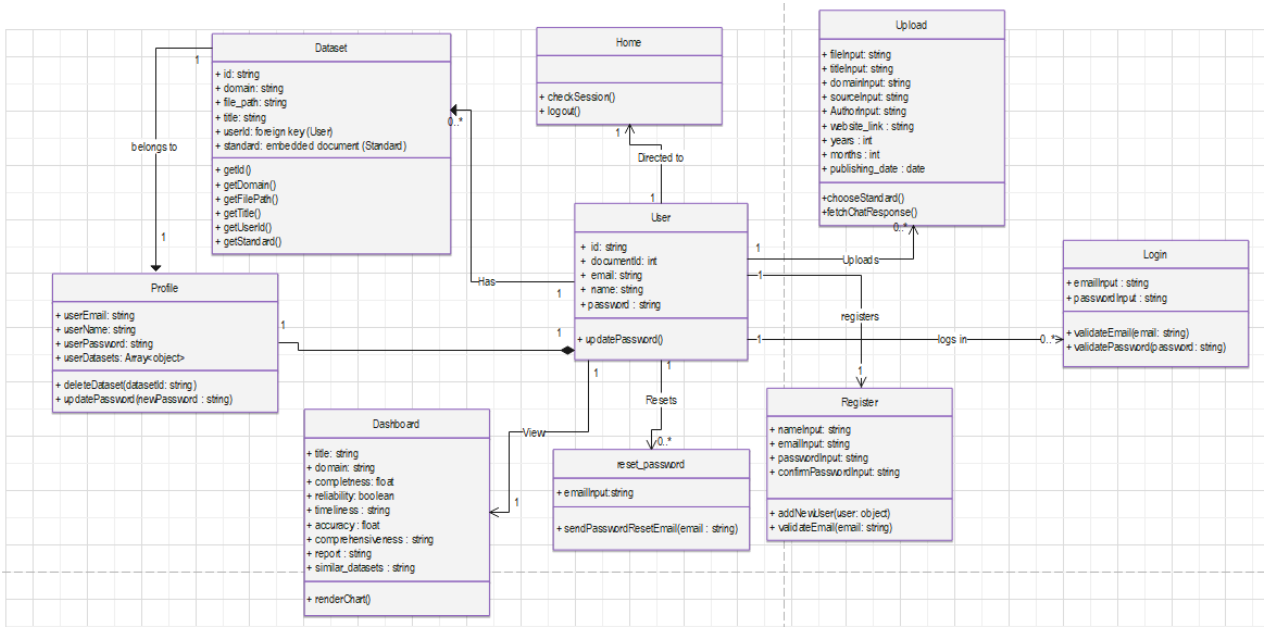


Figure 4 Class Diagram

4.3.3 Component Level Design

In this section we will provide the flowcharts and pseudocodes for our 5 major components.

4.3.3.1 Upload Feature Flowchart

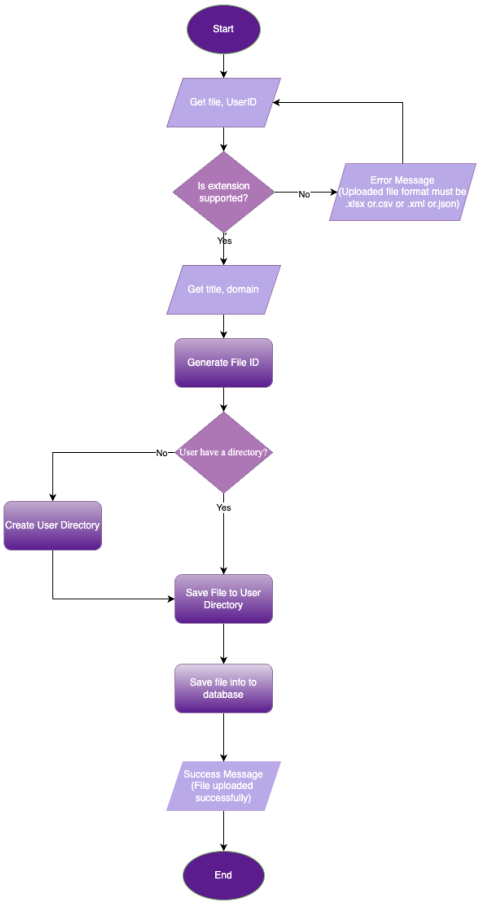


Figure 5 Uplolad feature's flow chart.

4.3.3.2 Completeness Measure Flowchart

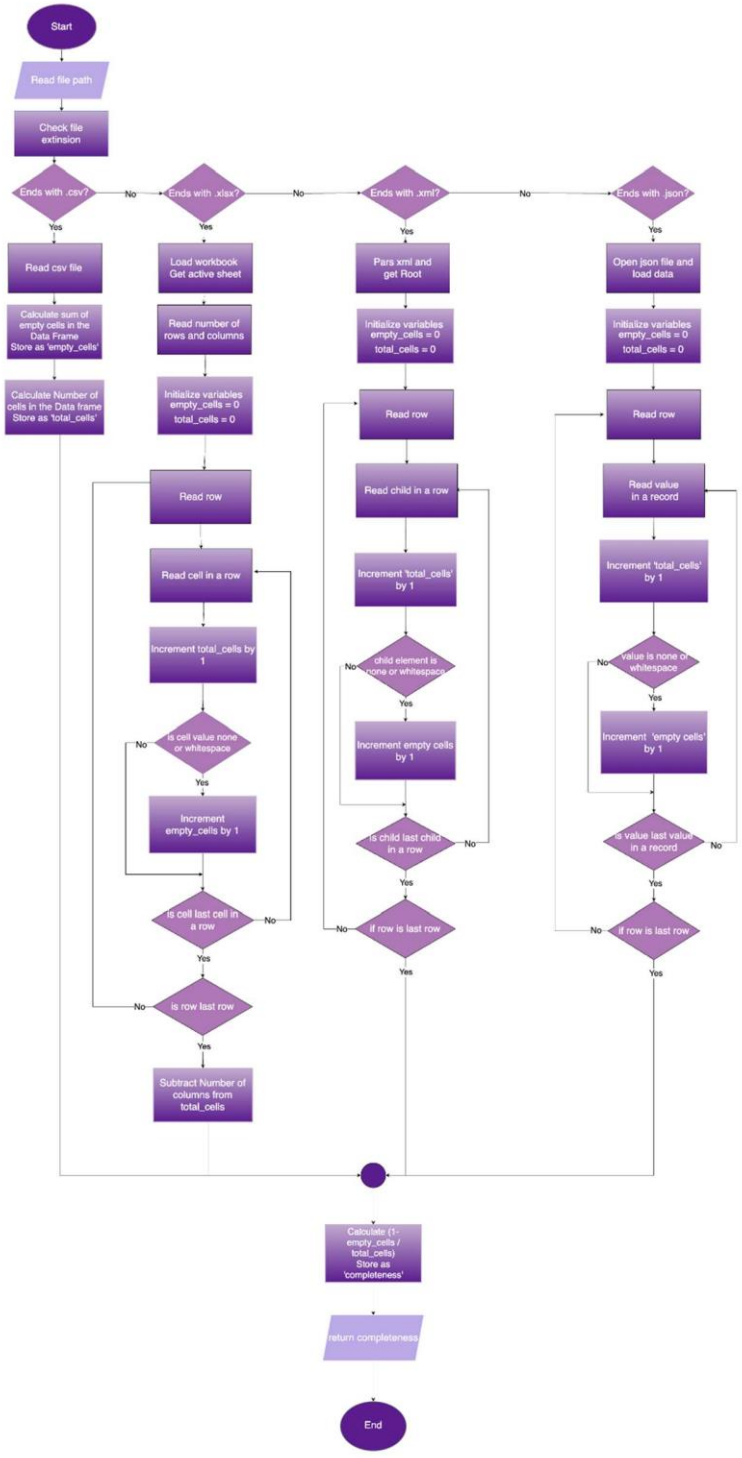


Figure 6 Completeness measure's flow chart.

4.3.3.3 Recommend Standards Pseudocode

```
BEGIN
    DISPLAY dropdown menu to the user
    READ selectedDomain from the dropdown menu

    IF selectedDomain is not empty:
        SET chatbotResponse by sending the prompt with selectedDomain to ChatGPT
        SPLIT chatbotResponse into sentences and assign them to sentences

        INITIALIZE an empty list called extractedStandards

        FOR each sentence in sentences:
            IF sentence contains any of the 6 standards (reliability, timeliness, accuracy, consistency, completeness, comprehensiveness):
                ADD the standards found in the sentence to extractedStandards
            ENDIF
        ENDFOR

        IF extractedStandards is not empty:
            SET response by concatenating the standards in extractedStandards
            DISPLAY response to the user
        ELSE:
            DISPLAY error message "No standards found in the response."
        ENDIF
    ELSE:
        DISPLAY error message "No domain selected."
    ENDIF
END
```

4.3.3.4 Timeliness Standard Pseudocode

```
1. BEGIN
2.     Display checkboxes for the user to choose "TIMELINESS"
3.     IF "TIMELINESS" is selected THEN
4.         Display input field for the user to enter the WEBSITE_LINK
5.         Display input fields for the user to enter YEARS and MONTHS
6.
7.         IF the WEBSITE_LINK contains "DATA.KSU.EDU.SA" THEN
8.             Perform web scraping to extract the publishing date from the WEBSITE_LINK
9.             Calculate the difference between the current date and the publishing date
10.        ELSE
11.            Display a date field for the user to manually provide the publishing date
12.            IF PROVIDED_PUBLISHING_DATE is provided THEN
13.                Convert PROVIDED_PUBLISHING_DATE to a datetime object
14.                Calculate the difference between the current date and the provided publishing date
15.            ELSE
16.                Set RESPONSE as "NO PUBLISHING DATE PROVIDED"
17.                RETURN RESPONSE
18.            ENDIF
19.        ENDIF
20.    ENDIF
```

4.3.3.5 Report generation feature

1. BEGIN
2. Initialize an empty array called results
3.
4. IF the "Generate Report" button is clicked THEN
5. Read the values of the chosen standards from user input and store them in the results array
6.
7. IF results is not empty THEN
8. Create an empty string called prompt
9. FOR each standard in results:
10. Append "Standard: " + standard + ", Result: " + results[standard] + "\n" to prompt
11. ENDFOR
12.
13. Send the prompt to the chatbot model
14.
15. Extract the generated report from the chatbot response
16.
17. Display the report to the user or perform further actions
18. ELSE
19. Set message as "No standards were chosen."
20. Display the message to the user
21. ENDIF
22. ENDIF
23. END

4.4 Data Design

4.4.1 Data Models

4.4.1.1 ER Diagram

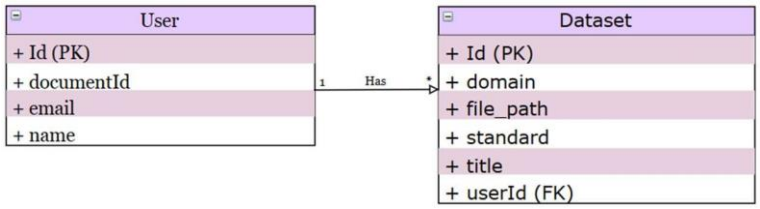


Figure 7 ER diagram

4.4.1.2 Non-relational Data Model

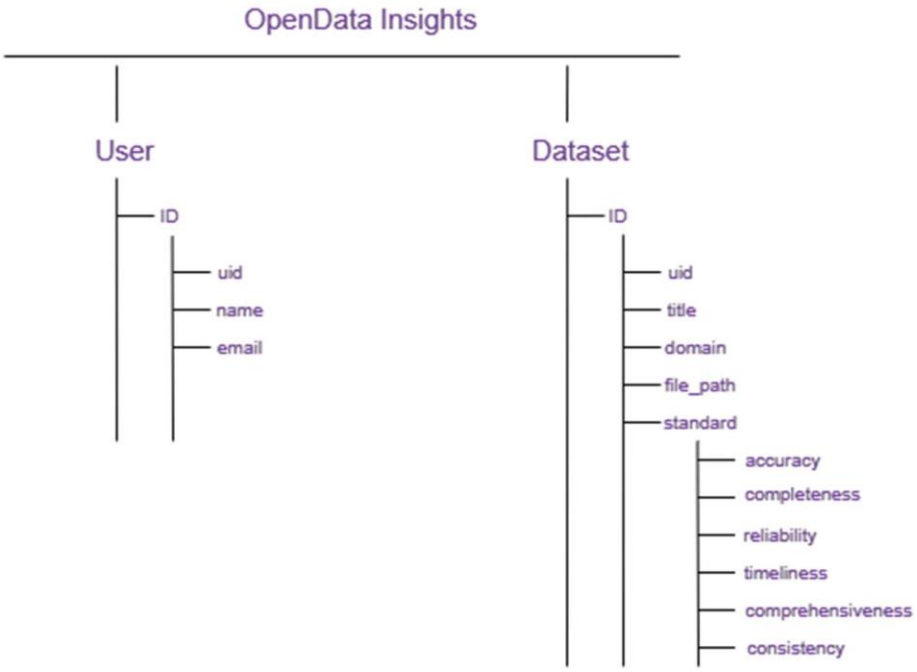


Figure 8 Hierarchical database model.

Collection: User

- Document: {userId}(string)

- Field: email (string)
- Field: name (string)

- Collection: Dataset
  - Document: {datasetId}(string)
    - Field: domain (string)
    - Field: file_path (string)
    - Field: standard (string)
    - Field: title (string)
    - Field: userId (string)

The relationship approach employed between the User and Dataset collections in our Firestore implementation is the referencing approach. This is demonstrated by the presence of the 'userId' field within the Dataset collection, which serves as a reference to the associated user in the User collection. By adopting referencing, we avoid redundant duplication of user information within each dataset document and instead store a reference to the corresponding user document. This approach optimizes storage efficiency, particularly in scenarios where multiple datasets are linked to a user.

4.4.2 Data Collection and Preparation

Our system is able to collect data from the users and apply the required preprocessing steps on it. The system can collect data in various formats such as CSV, XLSX, XML, or JSON. This feature allows users to provide their own data for analysis. Once uploaded, the dataset is stored in a permanent directory within the "static/files" directory, organized based on the user's ID. Python, being a commonly used language for data manipulation and analysis, plays a crucial role in our code.

To handle different file formats and perform data manipulation operations, we leverage several libraries. For example, we use pandas to read and manipulate CSV files, such as using pd.read_csv to load the data. With pandas, we can calculate the number of empty cells in a DataFrame using isnull().sum(), as well as determine the total number of cells by

subtracting the number of columns from the total size of the DataFrame (df.size - df.columns.size).

Similarly, we utilize openpyxl to work with Excel files. It allows us to read and manipulate these files using functions like openpyxl.load_workbook. By iterating over the rows and cells in the sheet, we can identify empty cells and calculate completeness metrics.

For JSON files, we employ the json library. It enables us to read the JSON data from the file using json.load(file). We iterate over the records and values in the data, counting empty cells and determining the completeness of the dataset.

In the case of XML files, we rely on the xml.etree.ElementTree library. It helps us parse and manipulate XML files, such as using ET.parse(path) to parse the XML and retrieve the root element (tree.getroot()). By iterating over the rows and child elements in the XML tree, we can identify empty cells and calculate completeness metrics.

4.5 Interface Design

To enhance user navigation and understanding of the system's organization, it is essential to utilize a site map that visually represents the structure and components. In this section, we provide the site map of "OpenData Insights".

Additionally, we outline the UX guidelines that were considered during the development of the website.

4.5.1 Site map

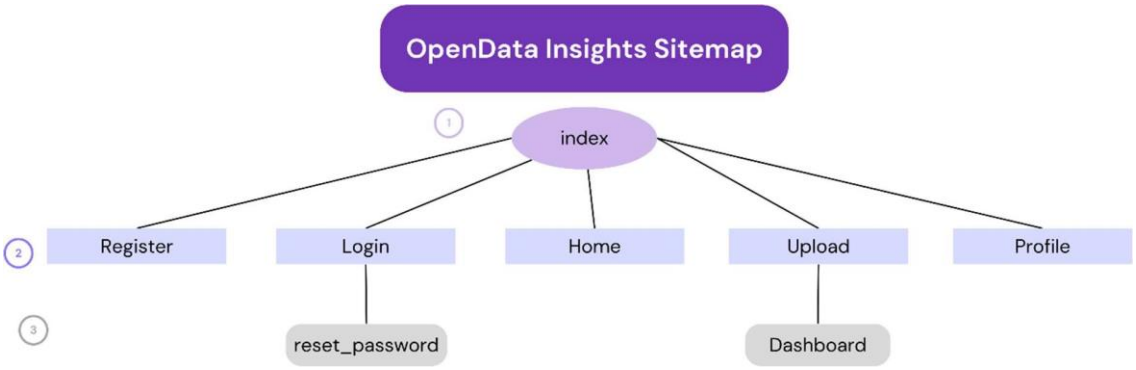


Figure 9 Site map.

4.5.2 UX guidelines

- **Familiarity:**

Familiarity plays a crucial role in website design as it enhances user experience by making users more comfortable and enabling them to interact with the website more quickly and effortlessly. This familiarity reduces the time users need to complete tasks compared to websites with unfamiliar styles. According to Jacobs Law, "Users spend most of their time on other sites. This means that users prefer your site to work the same way as all the other sites they already know" [28].

In our website, we have implemented this principle by incorporating familiar icons for each task (See Figure 10) and ensuring that the upload feature works and appears in a manner consistent with most websites worldwide (See Figure 11).

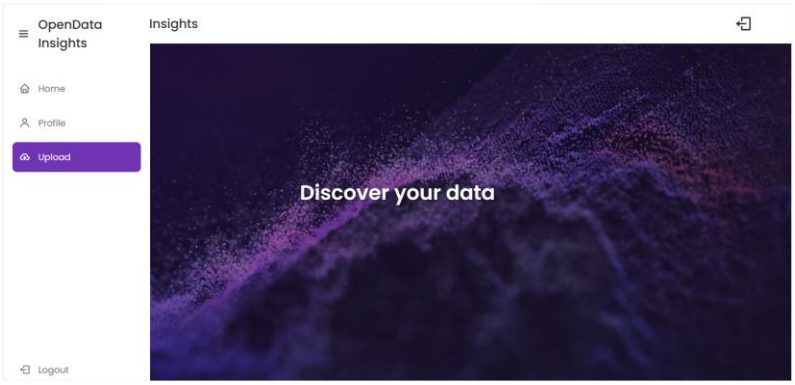


Figure 10 Sidebar Icons.

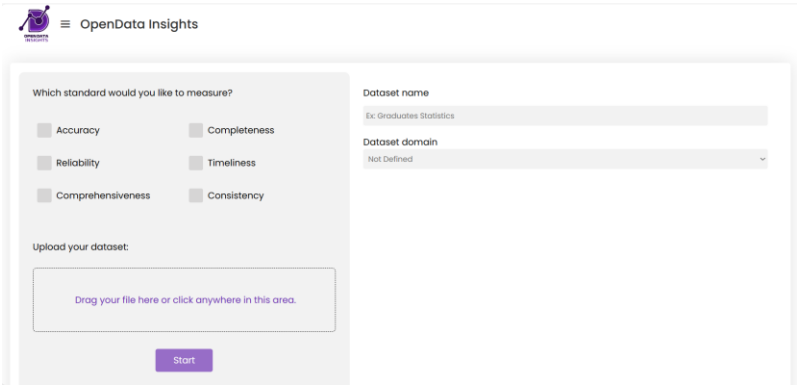


Figure 11 Upload Page.

- **Consistency:**

Consistency in design is essential for creating a comfortable user experience. It enables users to navigate and interact with various elements and pages of a website seamlessly. By maintaining consistency, we ensure that users can learn and adapt quickly and effortlessly. This leads to improved usability, higher user satisfaction, and increased engagement.

In our website, the sidebar navigation menu maintains a consistent style and position across all pages, ensuring that users can easily navigate without getting lost. Additionally, visual elements such as fonts, colors, and buttons remain consistent throughout the website, creating a visually cohesive and comfortable experience for users. These designs we have made contribute to a visually appealing and user-friendly interface.

- **Predictability:**

Predictability is essential in UX design. It refers to the ability of users to accurately anticipate the outcome of their interactions. It focuses on providing users with a sense of control and understanding, which allows them to confidently predict how the system will respond to their actions. When users can predict the outcome of their interactions, they can make informed decisions and navigate the interface more efficiently. This sense of predictability gives users a feeling of control and the ability to guess what will happen next, leading to a more intuitive and satisfying user experience.

In our website, we have applied predictability into the upload function. When a user selects a file from their device, they can anticipate that it has been successfully uploaded by observing the appearance of a file icon (See Figure 12 & 13). Also, when the "Start" button is in a light or disabled state as shown in Figure 14, users can predict that it is currently unable to perform any action, and when it became colored as shown in Figure 15, it means it is enabled now.

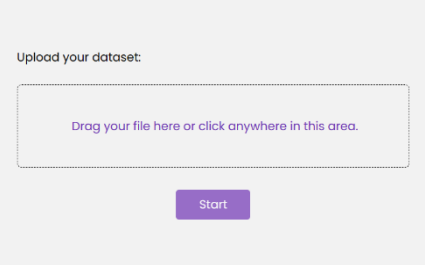


Figure 13 Upload section before uploading the file.

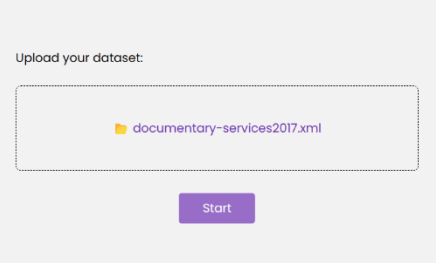


Figure 12 Upload section after uploading the file.

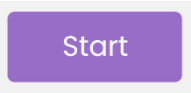


Figure 14 Disable Button.

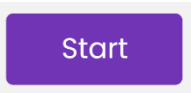


Figure 15 Enable Button.

- **Substitutivity:**

Substitutivity in user experience (UX) offers users alternative ways of specifying input or viewing output. It emphasizes providing flexibility and freedom for users to interact with a system or interface in different ways, based on their preferences, needs, or abilities. The aim is to provide alternative options or paths that can achieve similar outcomes or goals for the user.

In our website we have allowed the users to upload their dataset whether by selecting from their device or by dragging it in the right section -by saying "Drag your file here or click anywhere in this area." (See Figure 16)- aiding to permit alternative ways of entering data, this makes us apply the substitutivity principle.

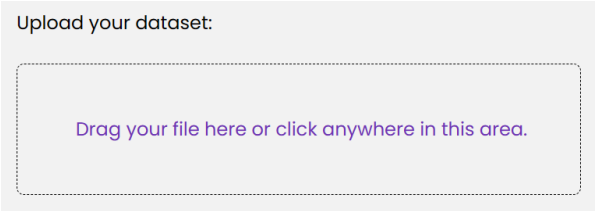


Figure 16 Showing alternative ways of entering data..

- **Observability (Browsability):**

Browsability in UX entails empowering users to examine the internal state of a system through the information presented in the interface. For instance, when a system is performing a time-consuming operation, it is important to display the current status of the operation, enabling users to stay informed about its progress.

To apply the browsability principle in our website, when the user clicks the "Start" button, they can know that the system has started processing their dataset and measuring the chosen quality, by the display of the "Processing..." text (See Figure 17). This feedback provides users with a clear understanding of the system's status and actions which allows them to investigate the system's internal state.

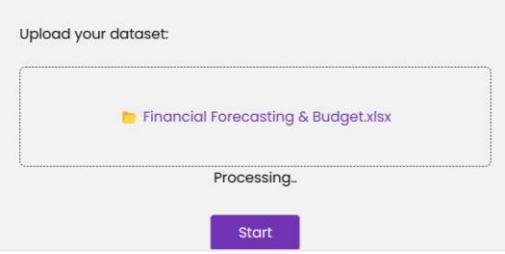


Figure 17 Display of "Processing..." text.

- **Recoverability:**

Recoverability is an important aspect of user experience design that focuses on minimizing user errors and providing clear error recovery paths. When users make mistakes or encounter errors, it is essential to provide them with accurate and concise error messages. These messages should clearly explain the nature of the problem allowing the users to resolve it correctly.

In our website, when the users enter a wrong password, invalid email format, or any other possible errors, the system will tell them specifically which error they made so they can solve it easily and not feel ambiguous (See Figure 18). By doing this we applied the recoverability principle.

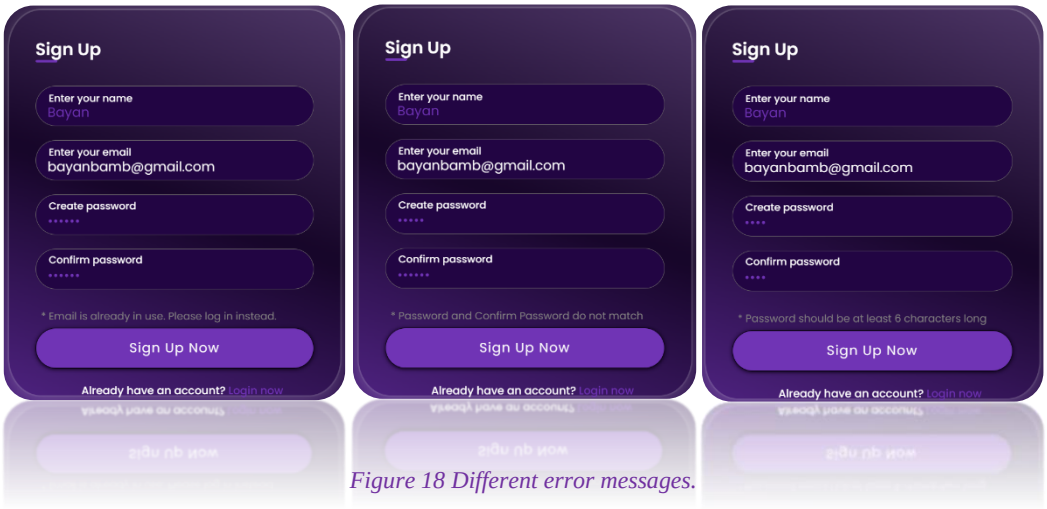


Figure 18 Different error messages.

4.6 System Implementation

4.6.1 Recommend standards:

The "Recommend Standards" feature in our system is designed to provide users with relevant standards based on the domain they have selected. The process begins by capturing the user's selected domain and sending it as input to ChatGPT. Using the OpenAI API, we leverage ChatGPT's natural language processing capabilities to generate a response that aligns with the user's domain. The response provided by ChatGPT is then carefully formatted to ensure clarity and readability. This formatting involves structuring the response to extract sentences that mention relevant standards. Once the response is formatted, it is displayed to the user in a user-friendly manner within the system's interface.

Following are the steps we used to implement “recommend standards”:

1. Import the required module which is openai from the OpenAI API
2. Set up the OpenAI API credentials by assigning our API key to openai.api_key.
3. Implement a function to interact with the OpenAI ChatGPT model.
    - ✓ Define a list of relevant standards.
    - ✓ Construct a prompt based on the provided domain and the list of standards.
    - ✓ Use the OpenAI API to make a completion request to the ChatGPT model.
    - ✓ Extract the response from the model's completion.
    - ✓ Call a function to extract suitable sentences.
        - Split the response into sentences.
        - Iterate over each sentence and each standard.

- If a standard is found in a sentence, add it to the set.
- If any suitable standards are found, return them.
- If no suitable standards are found, return an appropriate message indicating the absence of recommendations.

✓ Return the suitable sentence.

4. Define a function to handle the POST request.
   - ✓ Extract the domain.
   - ✓ Call the chat_with_gpt(domain) function to get the chat response.
   - ✓ If the chat response is not empty, format the response message.
   - ✓ If the chat response is empty or "null", set the response message to "No suitable standards were found".
   - ✓ Return the response message.

5. Return the response message as a JSON object using jsonify
6. Integrate the /fetch_chat_response endpoint with an AJAX request and display it to user.

```python
import openai
from urllib.parse import urlparse
from standards import CalcCompletness, CalcRelaiability
from flask import Flask, request, render_template, make_response,  redirect, url_for,
from flask_mail import Mail, Message
import firebase_admin
from firebase_admin import credentials, firestore, auth

app = Flask(__name__, template_folder='templates')

# Set up OpenAI API credentials
openai.api_key = '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'
```

```python
def chat_with_gpt(domain):
    standards = ['reliability', 'consistency', 'completeness', 'comprehensiveness', 'timeliness', 'accuracy']
    prompt = f"For a dataset in the {domain} domain, what are the most suitable standards out
    of {', '.join(standards)}?"

    # Set up the chatbot conversation
    chatbot_response = openai.completions.create(
        model="text-davinci-003",
        prompt=prompt,
        max_tokens=100,
        n=1,
        stop=None,
        temperature=0.7
    )

    chat_response = chatbot_response.choices[0].text.strip()

    suitable_sentence = extract_suitable_sentence(chat_response, standards)

    return suitable_sentence
```

```python
def extract_suitable_sentence(response, standards):
    # Split the response into sentences
    sentences = response.split('. ')

    # Look for unique sentences containing the suitable standards
    suitable_standards = set()
    for sentence in sentences:
        for standard in standards:
            if standard in sentence:
                suitable_standards.add(standard)

    if suitable_standards:
        return ', '.join(suitable_standards)
    else:
        return "No suitable standards were found in the chat response."

```

```python
#for chatGPT response (recommend)
@app.route('/fetch_chat_response', methods=['POST'])
def fetch_chat_response():
    domain = request.json.get('domain')
    chat_response = chat_with_gpt(domain)

    if chat_response and chat_response != "null":
        response_message = "We recommend you to choose these standards: " + chat_response
    else:
        response_message = "No suitable standards were found in the chat response."

    return response_message
```

```javascript
function fetchChatResponse() {
  const domainSelect = document.getElementById("domain");
  const selectedDomain = domainSelect.value;

  if (selectedDomain === "Not Defined") {
  const chatResponseElement = document.getElementById("chat-response");
  chatResponseElement.textContent = "Can not recommend because no domain is specified";
  return;
  }
  // Make an AJAX request to fetch the chat response
  const xhr = new XMLHttpRequest();
  xhr.open("POST", "/fetch_chat_response", true);
  xhr.setRequestHeader("Content-Type", "application/json");

  xhr.onreadystatechange = function () {
    if (xhr.readyState === XMLHttpRequest.DONE && xhr.status === 200) {
      const chatResponseElement = document.getElementById("chat-response");
      chatResponseElement.textContent = xhr.responseText;
    }
  };
  const data = JSON.stringify({ domain: selectedDomain });
  xhr.send(data);
}
```

Figure 19 Recommend standard code.

4.6.2 Upload file (dataset).

The "Upload dataset " function is a feature that enables users to upload and import open dataset into the system for them to assess.

The purpose of the "Upload dataset" function is to provide a convenient and user-friendly way to bring external open datasets into the system. Users can select a file from their device for the upload process.

After the file is successfully uploaded, the system undertakes validation procedures to ensure the file format aligns with the acceptable types (CSV, XLSX, XML, or JSON).

In order to implement the "Upload dataset" function, the following steps were followed:

- The code in Figure (20) below shows receiving file stream on user upload by checking if a file has been submitted in the request. If not or if the filename is empty, it returns to the upload page with a message indicating that no file is selected.

```
113 @app.route('/dashboard', methods=['POST'])
114 def dashboard():
115     if request.method == 'POST':
116         if 'file' not in request.files or request.files['file'].filename == '':
117             return render_template('upload.html', message='No file selected.')
118
119         file = request.files['file']
120
```

Figure 20 File upload.

- In the code shown in Figure (21), several steps are performed. Firstly, the file type is checked to determine if it matches the supported extensions (xlsx, csv, xml, and json). Next, a unique file ID is generated using the uuid.uuid4() function from the uuid module [29]. Following that, a directory is created for the user if it does not already exist. Finally, the uploaded file is saved in the user's directory with a filename constructed from the file ID and the original filename.

```
128
129         if file and allowed_file(file.filename):
130             fileId = str(uuid.uuid4())
131             user_directory = os.path.join("static", "files", userId)
132             if not os.path.exists(user_directory):
133                 os.makedirs(user_directory)
134             file_path = os.path.join(
135                 user_directory, f"{fileId}---{file.filename}")
136             file.save(file_path)
137
```

Figure 21 File Handling and Directory Creation.

- The code in Figure (22) below constructs dataset_data dictionary containing information about the uploaded file and dataset standards then save the dataset info to the firestore database.

```
191             dataset_data = {
192                 'file_path': file_path,
193                 'domain': domain,
194                 'title': title,
195                 'userId': userId,
196                 'standard': {
197                     'accuracy' : accuracy_val,
198                     'completness' : completeness_val,
199                     'reliability' : reliability_val,
200                     'timeliness' : timeliness_val,
201                     'comprehensiveness' : comprehensiveness_val,
202                     'consistency' : consistency_val,
203                 },
204                 'id' : fileId
205             }
206             collection = db.collection('dataset')
207             res = collection.add(dataset_data)
```

Figure 22 Saving Uploaded File Information to Firestore Database.

4.6.3 Evaluating the completeness standard.

The "calculateCompletenessStandard" function is designed to fulfill the user's requirement of assessing the completeness of an open dataset and providing a quality assessment result.

To implement the "completeness calculation" function, the following steps were followed:

- This function as shown in Figure (23) takes a file path and determines the file type by its extension. Then, it calls the corresponding function (CSV, XLSX, XML, or JSON) to calculate completeness based on the file type.

  Each function follows a similar pattern:

  - It initializes counters for empty cells (empty_cells) and total cells (total_cells).
  - It iterates through the relevant elements (rows, cells, or values) in the file format.
  - For each element, it checks if the content is empty (None or whitespace) and updates the counters accordingly.
  - After iterating through all elements, it calculates completeness using the formula: (1 - empty_cells / total_cells) * 100.
  - The completeness value is then returned.

```
8
9   def CalcCompletness(path):
10      if (path.endswith('.csv')):
11          return CSV(path)
12      if (path.endswith('.xlsx')):
13          return XLSX(path)
14      if (path.endswith('.xml')):
15          return XML(path)
16      if (path.endswith('.json')):
17          return JSON(path)
18
```

Figure 23 Evaluating the completeness.

- Figure (24) shows the function calculating completeness for CSV files. It uses the pandas library [30] to read the CSV file, count the number of empty cells, and calculate the completeness percentage.

```python
def CSV(path):
    df = pd.read_csv(path)
    empty_cells = df.isnull().sum().sum()

    total_cells = df.size

    completeness = (1 - empty_cells / total_cells) * 100

    print("*** CSV *** ", completeness)
    return completeness
```

Figure 24 Evaluating the completeness for CSV files.

- As shown in Figure (25) the function calculates completeness for Excel (XLSX) files. It uses the openpyxl library [31] to load the Excel workbook, iterates through all cells in each row, counts the number of empty cells, and calculates the completeness percentage.

```python

def XLSX(path):
    wb = openpyxl.load_workbook(path)
    sheet = wb.active

    num_rows = sheet.max_row
    num_cols = sheet.max_column

    empty_cells = 0
    total_cells = 0

    for row in sheet.iter_rows(min_row=1, max_row=num_rows, min_col=1, max_col=num_cols):
        for cell in row:
            total_cells += 1
            if cell.value is None or cell.value == "":
                empty_cells += 1

    total_cells -= num_cols
    completeness = (1 - empty_cells / total_cells) * 100
    print("*** XLSX *** ", completeness)
    return completeness

```

Figure 25 Evaluating the completeness for Excel (xlsx) files.

- As shown in Figure (26) the function calculates completeness for XML files. It uses the ElementTree module [32] to parse the XML file, iterates through the 'row' elements, counts the number of empty cells, and calculates the completeness percentage.

```python
def XML(path):
    tree = ET.parse(path)
    root = tree.getroot()

    empty_cells = 0
    total_cells = 0

    for row in root.iter('row'):
        for child in row:
            total_cells += 1
            if child.text is None or child.text.strip() == "":
                empty_cells += 1

    completeness = (1 - empty_cells / total_cells) * 100
    print("*** XML *** ", completeness)
    return completeness
```

Figure 26 Evaluating the completeness XML files.

- As shown in Figure (27) the function calculates completeness for JSON files. It reads the JSON file, iterates through the records and values, counts the number of empty cells, and calculates the completeness percentage.

```python
def JSON(path):
    data = None
    with open(path, 'r', encoding='utf-8') as file:
        data = json.load(file)

    empty_cells = 0
    total_cells = 0

    for record in data:
        for value in record.values():
            total_cells += 1
            if value is None or (isinstance(value, str) and value.strip() == ""):
                empty_cells += 1

    completeness = (1 - empty_cells / total_cells) * 100
    print("*** JSON *** ", completeness)
    return completeness
```

Figure 27 Evaluating the completeness JSON files.

4.6.4 Evaluating the Timeliness

The “timeliness” function seeks to determine the uploaded dataset's "up-to-dateness," whether it is "up-to-date" or "out-of-date," and the difference in months or years. The user will be prompted to enter the URL of the website from which he obtained the data. Our system will then verify whether the data came from, say, "data.ksu.edu," and if so, the user won't have to enter the publication date manually because the system will extract it using web scraping techniques.

- User Input and Displaying Timeliness Fields:
  - Users input website details including the link, years, and optionally the publishing date as shown in the following Figure (28).

```
<div id="timelinessQSection">
                            <div class="input-container">
                                <label>Website Link:</label>
                                <input name="website_link" id="website_link" type="text"
                                    placeholder="Ex: data.ksu.edu.sa">
                            </div>

                            <div class="input-container">
                                <label style="margin-bottom: 10px;">Number of Years and Months difference:</label>
                                <div class="flexbox">
                                    <div class="input-container">
                                        <label for="years">Years:</label>
                                        <input name="years" id="years" type="number" placeholder="Ex: 5">
                                    </div>
                                    <div class="input-container">
                                        <label for="months">Months:</label>
                                        <input name="months" id="months" type="number" placeholder="Ex: 3">
                                    </div>
                                </div>
                            </div>


                            <div class="input-container">
                                <label for="publishing_date" id="publishing_date_label_visible">Publishing Date:</label>
                                <input type="date" id="publishing_date" name="publishing_date">
                            </div>

                        </div>
```

Figure 28 Displaying Timeliness Fields.

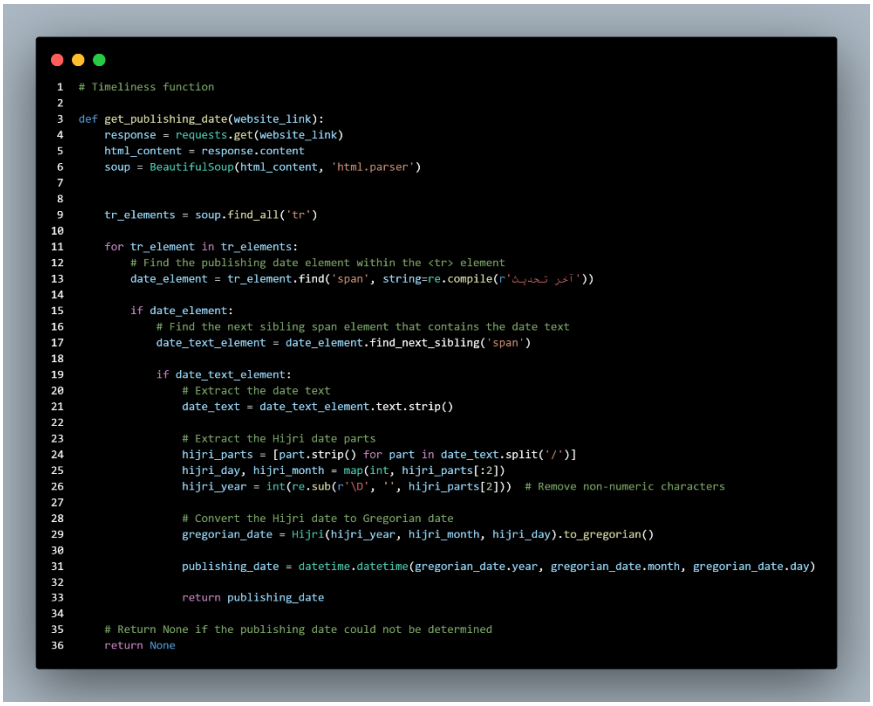
  - The interface dynamically adjusts to show or hide timeliness fields based on user selection of timeliness checkbox as show in Figure (29).

```

//For timeliness section
function toggleTimelinessQ() {
    var timelinessCheckbox = document.getElementById("Timeliness");
    var timelinessQSection = document.getElementById("timelinessQSection");
    var websiteLinkInput = document.getElementById("website_link");
    var yearsInput = document.getElementById("years");
    var publishingDateLabel = document.getElementById("publishing_date_label_visible");
    var publishingDateInput = document.getElementById("publishing_date");

    if (timelinessCheckbox.checked) {
        timelinessQSection.style.display = "block";
        websiteLinkInput.required = true;
        yearsInput.required = true;

        websiteLinkInput.addEventListener("change", function () {
            if (!websiteLinkInput.value.includes("data.gov.sa") && !websiteLinkInput.value.includes("data.ksu.edu.sa")) {
                publishingDateLabel.style.display = "inline-block";
                publishingDateInput.style.display = "inline-block";
                publishingDateInput.required = true;
            } else {
                publishingDateLabel.style.display = "none";
                publishingDateInput.style.display = "none";
                publishingDateInput.required = false;
            }
        });
    } else {
        timelinessQSection.style.display = "none";
        websiteLinkInput.required = false;
        yearsInput.required = false;
        publishingDateLabel.style.display = "none";
        publishingDateInput.style.display = "none";
        publishingDateInput.required = false;
    }
}
```

Figure 29 Dynamic timeliness fields.

- Checking the Link for Web Scraping:
  - If the link corresponds to specified data portals ("data.gov.sa" or "data.ksu.edu.sa"):
    - Web scraping is triggered to extract the publishing date from the website as shown in Figure (30).

- The publishing date is retrieved and used for timeliness calculation.



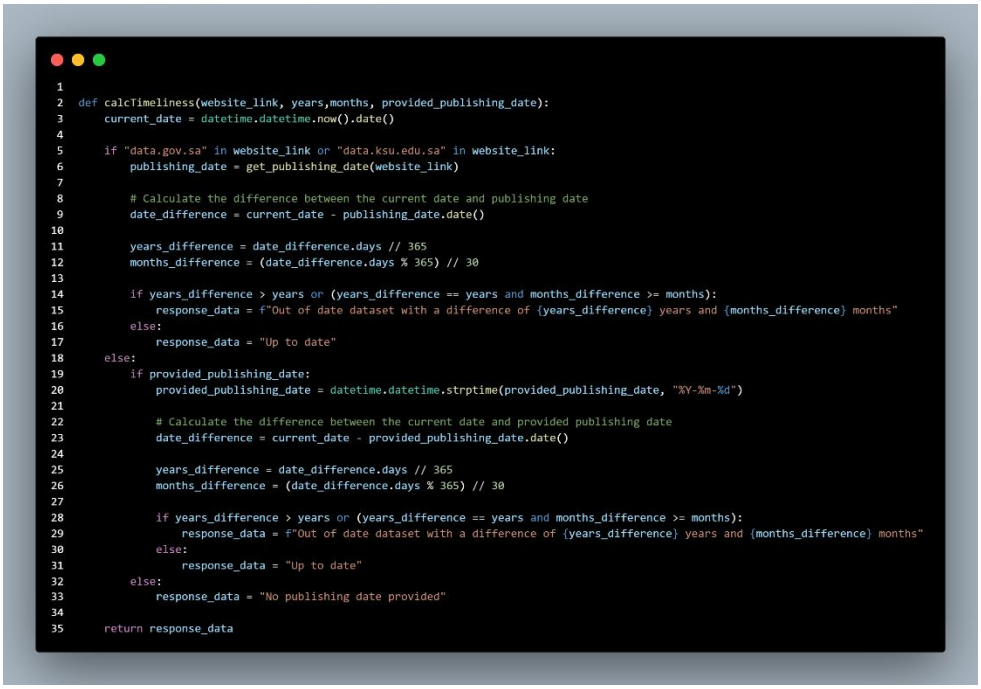
```python
# Timeliness function

def get_publishing_date(website_link):
    response = requests.get(website_link)
    html_content = response.content
    soup = BeautifulSoup(html_content, 'html.parser')


    tr_elements = soup.find_all('tr')

    for tr_element in tr_elements:
        # Find the publishing date element within the <tr> element
        date_element = tr_element.find('span', string=re.compile(r'آخر تحديث'))

        if date_element:
            # Find the next sibling span element that contains the date text
            date_text_element = date_element.find_next_sibling('span')

            if date_text_element:
                # Extract the date text
                date_text = date_text_element.text.strip()

                # Extract the Hijri date parts
                hijri_parts = [part.strip() for part in date_text.split('/')]
                hijri_day, hijri_month = map(int, hijri_parts[:2])
                hijri_year = int(re.sub(r'\D', '', hijri_parts[2]))  # Remove non-numeric characters

                # Convert the Hijri date to Gregorian date
                gregorian_date = Hijri(hijri_year, hijri_month, hijri_day).to_gregorian()

                publishing_date = datetime.datetime(gregorian_date.year, gregorian_date.month, gregorian_date.day)

                return publishing_date

    # Return None if the publishing date could not be determined
    return None
```

Figure 30 Web scraping.

- If the link doesn't match the specified data portals:
  - The provided publishing date (if any) is used for timeliness calculation.

➢ Calculate Timeliness:

- Timeliness is calculated based on the difference between the current date and the publishing date.
- If the dataset is older than the specified threshold (years and months), it's flagged as out of date.
- Otherwise, it's considered up to date as shown in Figure (31).



```python

def calcTimeliness(website_link, years,months, provided_publishing_date):
    current_date = datetime.datetime.now().date()

    if "data.gov.sa" in website_link or "data.ksu.edu.sa" in website_link:
        publishing_date = get_publishing_date(website_link)

        # Calculate the difference between the current date and publishing date
        date_difference = current_date - publishing_date.date()

        years_difference = date_difference.days // 365
        months_difference = (date_difference.days % 365) // 30

        if years_difference > years or (years_difference == years and months_difference >= months):
            response_data = f"Out of date dataset with a difference of {years_difference} years and {months_difference} months"
        else:
            response_data = "Up to date"
    else:
        if provided_publishing_date:
            provided_publishing_date = datetime.datetime.strptime(provided_publishing_date, "%Y-%m-%d")

            # Calculate the difference between the current date and provided publishing date
            date_difference = current_date - provided_publishing_date.date()

            years_difference = date_difference.days // 365
            months_difference = (date_difference.days % 365) // 30

            if years_difference > years or (years_difference == years and months_difference >= months):
                response_data = f"Out of date dataset with a difference of {years_difference} years and {months_difference} months"
            else:
                response_data = "Up to date"
        else:
            response_data = "No publishing date provided"

    return response_data
```

Figure 31 Timeliness function.

4.7 GitHub link

GitHub: https://github.com/OpenDataInsight/2023-GP1-7-Final-Release-1.git

5. System Evaluation

5.1 Experimental Results

The result in the following Table 6,7,8,9 shows that the system performs well in computing data quality standards for different formats using 20 different datasets. The tables show the performance of computing the completeness, accuracy, consistency, and comprehensiveness standards and compare the system output and the manual output by humans. The reason for this comparison is to make sure that the system computed the required standard correctly regardless of the data format.

Table 6 Dataset Completeness Assessment Results

Dataset No.	Dataset Extension	Is Complete?	No. Incomplete Cells	No. Complete Cells	No. Of Cells	Manual Result	System Result	Passed the test?
1	**CSV**	Yes	0	15	15	100%	100%	✓
2	**JSON**	Yes	0	15	15	100%	100%	✓
3	**XLSX**	Yes	0	15	15	100%	100%	✓
4	**XML**	Yes	0	15	15	100%	100%	✓
5	**CSV**	No	2	13	15	86.66%	86.66%	✓
6	**JSON**	No	3	12	15	80%	80%	✓
7	**XLSX**	No	1	14	15	93.33%	93.33%	✓
8	**XML**	No	4	11	15	73.33%	73.33%	✓
9	**XLSX**	Yes	0	16	16	100%	100%	✓
10	**XLSX**	No	3	13	16	83.33%	83.33%	✓

Table 7 Dataset Accuracy Assessment Results

Dataset No.	Dataset Extension	Is accurate?	Manual Result	System Result	Passed the test?
1	CSV	Yes	100%	94.69%	✓
2	JSON	Yes	100%	95.83%	✓
3	XML	Yes	100%	97.83%	✓
4	XLSX	Yes	100%	91.59	✓

Table 8 Dataset Comprehensiveness Assessment Results.

Dataset No.	Dataset Extension	Is Comprehensive?	Manual Result	System Result	Passed the test?
1	CSV	Yes	Yes	Yes	✓
2	JSON	No	No	No	✓
3	XLSX	Yes	Yes	Yes	✓
4	XML	No	No	No	✓

Table 9 Dataset Consistency Assessment Results.

Dataset No.	Dataset Extension	Manual Result (TYPE CONSIST	Manual Result (LANGUAGE	System Result (TYPE CONSIST	System Result Result (LANGUAGE	Passed the test?

		ENT?)	CONSISTENT?)	ENT?)	CONSISTENT?)	
1	**JSON**	Yes	Yes	Yes	Yes	✓
2	**CSV**	No	No	No	No	✓
3	**XML**	Yes	Yes	Yes	Yes	✓
4	**XLSX**	Yes	No	Yes	No	✓
5	**XLSX**	No	No	No	No	✓

5.2 User Acceptance Testing

In this section, we provide a summary of the User Acceptance Testing (UAT) that was conducted on the system. The UAT aimed to evaluate the system's acceptance by its intended users. For that we have tested 20 users, 10 publishers and 10 consumers with varying technical expertise. During the testing, the users were presented with specific scenarios to complete, while our team carefully observed their behaviors and gathered comments from users while testing. Following the testing phase, the users were provided with a questionnaire form to further express their insights, which provided us with additional valuable information regarding the system's ability to meet its intended objectives.

The testing results were extensively analyzed to determine the system's ability to meet its intended objectives. In addition, they were used to identify areas for improvement.

It turned out that the UAT was a valuable tool for evaluating the user experience. It ensured that the system does what it was designed to do, and that it meets the needs of its intended users. The insights gained from the testing process serve as a foundation for ongoing system enhancements.

5.2.1 Demographics of Participants

Table 10 Demographics of Participants.

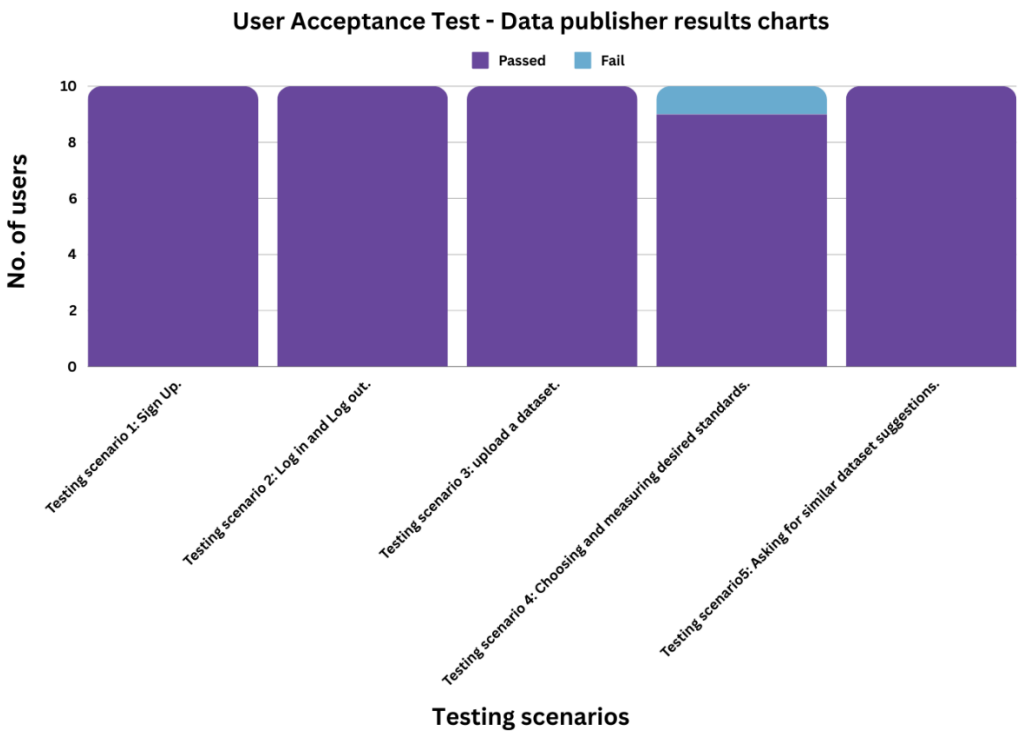
Participant No	Age	Gender	Technical background	User Type
1	*37*	Female.	Has technical background.	Data publisher.
2	*36*	Female.	Has technical background.	Data publisher.
3	*45*	Female.	Has technical background.	Data publisher.
4	*40*	Female.	Has technical background.	Data publisher.
5	*28*	Female.	Has technical background.	Data publisher.
6	*25*	*Male.*	Has technical background.	Data publisher.
7	*23*	*Male.*	Has technical background.	Data publisher.
8	*23*	Female.	Has technical background.	Data publisher.
9	*25*	Female.	Has technical background.	Data publisher.
10	*25*	Female.	Has technical background.	Data publisher.
11	25	Female.	does not have.	Data consumer.
12	21	Male.	does not have.	Data consumer.
13	27	Female.	Has Technical Background.	Data consumer.
14	24	Female.	Has Technical Background.	Data consumer.
15	19	Female.	does not have.	Data consumer.
16	21	Female.	Has Technical Background.	Data consumer.
17	35	Male.	Has Technical Background.	Data consumer.

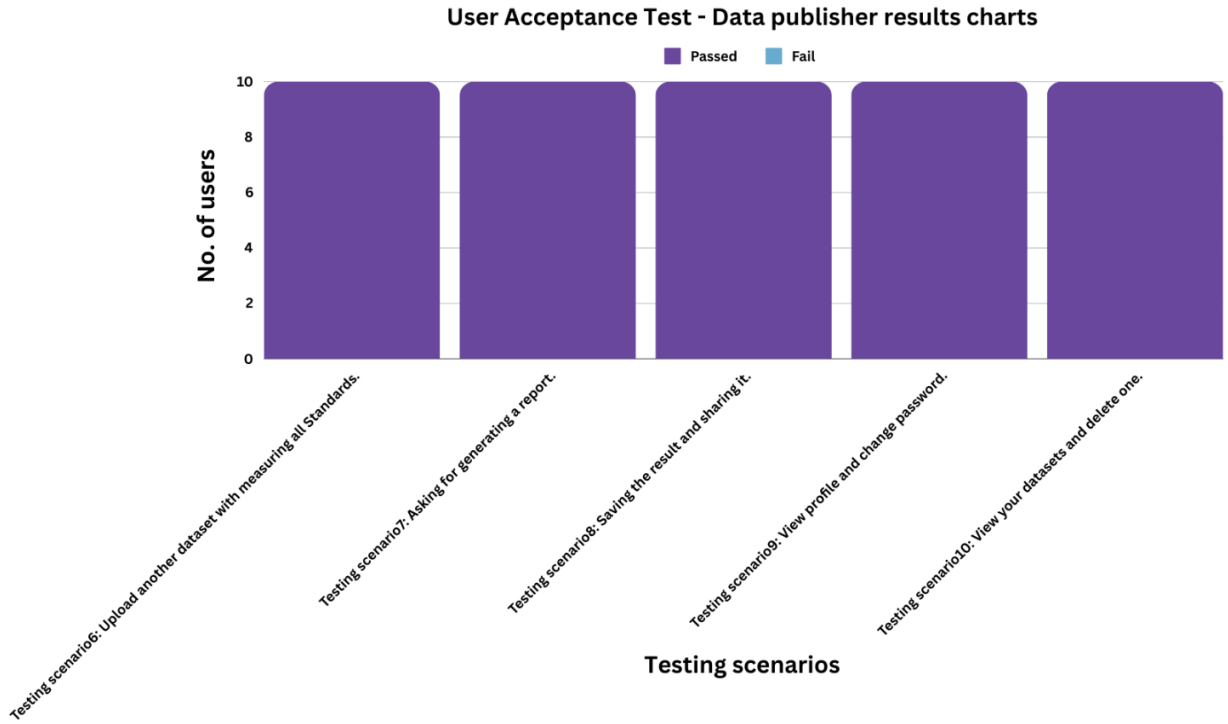
18	32	Female.	Has Technical Background.	Data consumer.
19	31	Female.	does not have.	Data consumer.
20	36	Female.	does not have.	Data consumer.

5.2.2 Questionnaire/Interview Results

In this section, we provide the findings from the conducted user testing. By offering a comprehensive overview of users' experiences and opinions these results serve as a valuable resource to inform future improvements and ensure that the system effectively meets their needs.

The following scenarios as shown in Figure (27) below have been observed by us as they are being tested by users. We determined whether the user successfully completed the features or encountered any issues by seeing how they used “openData insights”, interacted with the interfaces, and reacted to the features.


User Acceptance Test - Data publisher results charts
Passed
Fail
No. of users
0
2
4
6
8
10
Testing scenario 1: Sign Up.
Testing scenario 2: Log in and Log out.
Testing scenario 3: upload a dataset.
Testing scenario 4: Choosing and measuring desired standards.
Testing scenario5: Asking for similar dataset suggestions.
Testing scenarios


User Acceptance Test - Data publisher results charts
Passed
Fail
No. of users
0
2
4
6
8
10
Testing scenario6: Upload another dataset with measuring all Standards.
Testing scenario7: Asking for generating a report.
Testing scenario8: Saving the result and sharing it.
Testing scenario9: View profile and change password.
Testing scenario10: View your datasets and delete one.
Testing scenarios

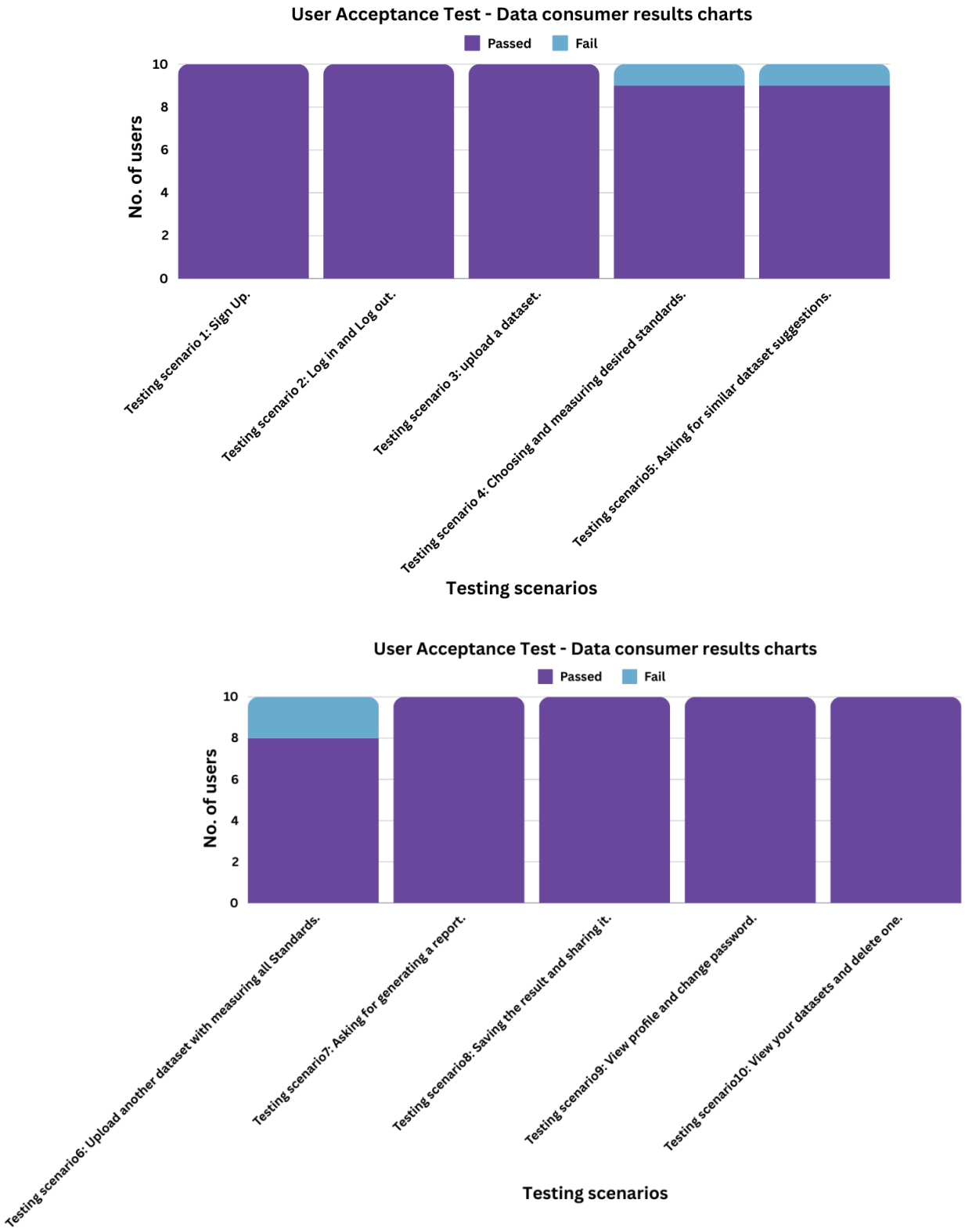


Figure 27 End users result chart.

According to the chart depicted in Figure (27), it can be deduced that all users have effectively completed the scenarios. Nevertheless, participant No.4, whose type is a publisher and 11, whose type is a consumer, encountered challenges in choosing and measuring desired standards scenario, as they were unaware of these standards and their meanings. However, after receiving an explanation of the standards, they were able to successfully pass the scenario. Additionally, participant No.12, whose type is a consumer, faced difficulties with "asking for similar dataset suggestions" and participants No.11,12 with "uploading another dataset with measuring all Standards".

The details of this questionnaire are listed in **APPENDIX G**: End user's questionnaire.

The results of the User Acceptance Testing questionnaire regarding open data publishers indicate positive feedback from the participants. Firstly, 16 participants identified as female and 4 as male. In terms of technical background, 15 participants answered yes and 5 answered no. The participants unanimously agreed that the dashboard's user interface was visually appealing and user-friendly. They also agreed that the charts (results) were effective and easy to understand. Additionally, all participants found the system easy to use and navigate except participants No.13,16 and 17 were Natural. All participants were able to interpret and understand the completeness result, and they found the view profile page clear and easy to understand. However, opinions differed when it came to understanding the consistency result, with 12 participants agreeing that they were able to interpret and understand the result, while 8 participants were neutral. All participants agreed that they were able to understand the report except participant No.15 was Natural. All participants agreed that checking the reliability of the dataset was intuitive and easy to understand. Regarding the overall user experience of the dataset upload function, 10 participants rated it as excellent, while 10 participants rated it as very good. Finally, 11 participants strongly agreed that they were satisfied with the system overall while 9 agreed. These positive responses indicate that the "OpenData Insights" system was well-received by the participants, highlighting the strengths of the system while also indicating areas that may require further attention and improvement.

5.3 Quality Attributes (NFR testing)

Table 11 NFR testing.

User story	Quality Attribute	Measure	Results
As a user, I want the system to be fast in response, within 10 seconds, so that I can quickly assess the quality of the open dataset without experiencing delays.	Performance: How responsive are the parts of the system as a whole?	**Load time:** It is a measurement of the time it takes for the browser to download and process all of the resources needed to render a webpage, such as scripts, images, and stylesheets. **Time To First Byte:** (TTFB) is the amount of time it takes for the browser to receive the first byte of data from the server after a request is made. It's a measure of how long it takes for the server to start processing and sending data in response to a request.	To begin, we utilized the GTmetrix website, where we input the URL of OpenData Insights into the designated text box. From there, we chose a test location from the provided drop-down menu. Initiating the test was as simple as clicking on the "Start test" button. GTmetrix then proceeded to assess the website's performance and generated a comprehensive report, which included the crucial metric of Page Load Time. After conducting the task three times, we determined that the maximum load time recorded was 1.25 seconds, as illustrated in Figure (28). Additionally, we examined the Time To

			First Byte (TTFB) using SEO Site Checkup. By inputting our website's link, we obtained a TTFB value of 0.009 seconds, falling within the acceptable green range, as depicted in Figure (31).
As a user, I want to intuitively understand and use the system within 10 mins without requiring special training so that I don't waste time learning and make any mistakes.	Usability: describes how simple a system is to use, how well users can complete particular activities, and how satisfied they are with the overall experience.	**Task completion rate** is commonly computed as the proportion of users who successfully finish a given task relative to the total number of users that attempted it.	We tested a feature that enables users to measure all six standards of a dataset after being briefed on the scenario. Six participants took part in the test, and each attempted to measure the standards. All participants successfully completed the task, resulting in a 100% feature success rate. This rate is calculated by dividing the number of users who successfully completed the task by the total number of users (6/6) and multiplying by 100,

			yielding 100% which was our target.
As a user, I want the system to be reliable 95% of the time when the user numbers increase so that I can access and use the system without degradation or downtime.	Scalability: describes a system that can adapt to increased load without compromising functionality, performance, or user experience.	**Response Time (in milliseconds)**: This measure tells us how quickly our website responds to user requests. **Throughput (requests per second)**: This measure tells us how many requests our website can handle in one second.	We employed Locust to evaluate our website's performance under increased concurrent user activity. With 20 users, we observed a median response time of 160 milliseconds as shown in Figure (32), indicating efficient handling of most requests, thus ensuring a smooth user experience. Although occasional delays were noted in the 95th and 99th percentiles, their impact on overall user satisfaction was minimal. Additionally, our current throughput of 6.7 requests per second as shown in Figure (32) effectively manages our current workload.

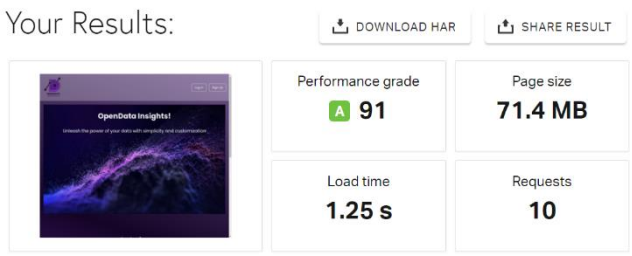


Figure 28 Performance test 1.

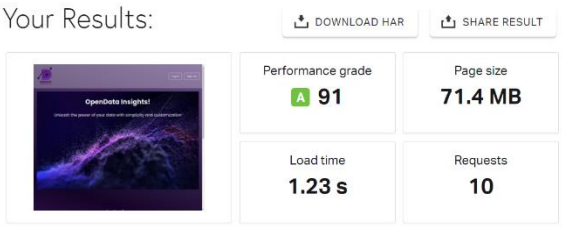


Figure 29 Performance test 2.

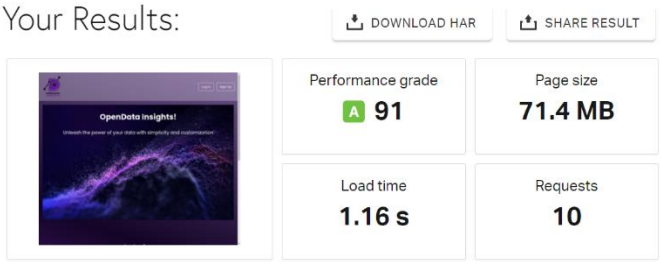


Figure 30 Performance test 3.

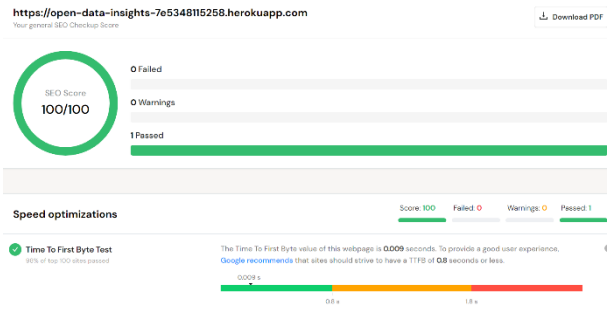
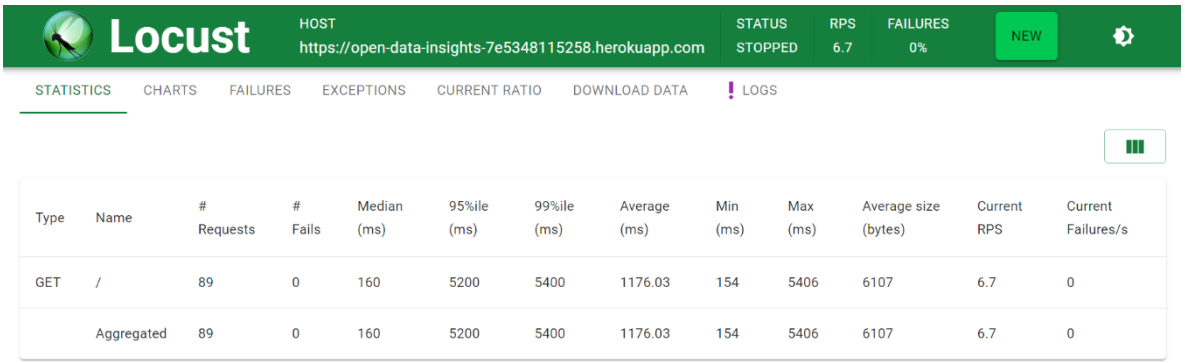


Figure 31 Performance TTFB.



Type	Name	# Requests	# Fails	Median (ms)	95%ile (ms)	99%ile (ms)	Average (ms)	Min (ms)	Max (ms)	Average size (bytes)	Current RPS	Current Failures/s
GET	/	89	0	160	5200	5400	1176.03	154	5406	6107	6.7	0
	Aggregated	89	0	160	5200	5400	1176.03	154	5406	6107	6.7	0

Figure 32 Scalability Test.

5.4 Discussion

The results obtained from the system evaluation for both open data publishers and consumers were generally positive. Most of the participants, regardless of their technical background, found the dashboard's user interface visually appealing and user-friendly. They also agreed that the charts were effective and easy to understand, were able to understand the generated report, and they found the process of checking the reliability standard intuitive and easy to understand. The completeness result and the view profile page were unanimously considered

clear and easy to understand by all participants.

In terms of comprehending the consistency result, it is worth noting that the majority of participants found it understandable, while four participants found it to be a natural concept. One participant, participant No. 19, who identified as a consumer, initially found it natural but required additional clarification due to being unfamiliar with quantitative and qualitative concepts. However, once we provided explanations for these concepts, participant No. 19 was able to interpret and grasp the result effectively. Furthermore, it is important to highlight that the lack of a % symbol in the chart depicting the quantitative columns created confusion for the remaining three participants who were publishers. This oversight hindered their ability to clearly interpret the chart. Based on this feedback, it is evident that incorporating the % symbol in the chart is necessary to alleviate any ambiguity and ensure a more intuitive understanding of the data.

For the system to be easy to use and navigate. It was observed that three participants, two publishers and a consumer, naturally responded to the system. This feedback can be attributed to the small sections present on the homepage which explains the system's major features, the participants assumed that these sections were clickable and would directly navigate them to the corresponding major features. Based on the results obtained, we have identified two potential enhancement options to improve the user experience. These options revolve around the small sections on the home page that explain the system's major features. One enhancement option is to make the small sections clickable, allowing users to navigate directly to the corresponding major features. Alternatively, we can enhance the sections by redesigning them to ensure greater clarity and provide comprehensive explanations of the system's major features. This approach aims to address any potential confusion or ambiguity that users might encounter. By presenting clear and concise information, we can help users better understand the system's functionalities and benefits. To determine the most effective enhancement option, it would be beneficial to consider factors such as user preferences, system requirements, and the overall user interface design.

When checking the user experience of the open dataset upload function and whether the users are satisfied with the system or not. Encouragingly, the responses were overwhelmingly positive, with participants rating the system as excellent or very good. This indicates a generally positive user experience. However, to gain deeper insights and delve further into their experiences and satisfaction, we also invited participants to share their comments at the end of the testing

session. Here are the comments provided by the participants:

1- The compare chart would be more understandable if the results of accuracy and completeness would be represented in different colors.
2- It would be better to use the garbage icon rather than the "x" to delete a dataset.
3- The upload date should be added to the open dataset information along with the other information so that the user knows when it was uploaded.
4-The text indicating the user's password was successfully changed should be presented in a clearer manner like a pop-up window.
5- In addition to suggesting suitable standards to users, the system should explain why the standards are suggested.
6- “Rephrase” would be a better name for the regenerate button.
7- It would be better if dataset suggestions always appeared without having to press a button.
8- It would be better if the report always appeared without having to press a button.

The above-mentioned comments that we collected from the users are valuable observations that will assist us in making significant improvements to the system, making it more suitable and user-friendly. We will work on incorporating these notes into our enhancements. However, we also recognize that some of these comments revolve around differences in people's preferences regarding the user interface such as using “Rephrase” instead of “Regenerate”, and making the suggestions and report appear without the need for pressing a button. Moreover, we think that these preferences are attributed to the participants themselves.

In conclusion, most participants regardless of their type and technical background have liked the system. They think it saves time and effort, also important and useful to anyone with open data. They also think that the covered standards are great.

When we tested our non-functional requirements, we focused on three main areas. First up was performance. We used GTmetrix to check load time and SEO Site Checkup for Time To First Byte (TTFB). Turns out, our maximum load time was just 1.25 seconds, well below the 10-second target. That tells us our system responds pretty fast. And for TTFB, it was just 0.009 seconds, showing the server processes data efficiently.

Next, we looked at usability. We used Task Completion Rate and all six participants easily finished the task of measuring dataset standards. That's a 100% success rate, indicating our system is easy to understand and use.

Lastly, we explored scalability. Using Locust, we found that even with 20 users at once, our median response time was just 160 milliseconds, showing we can handle user requests well. Though there were occasional delays in the 95th and 99th percentiles, they didn't affect user satisfaction much. Plus, our throughput of 6.7 requests per second shows we can handle increased user activity without any issues.

6. Conclusions and Future Work

6.1 Global and local impact.

The development of "OpenData Insights" is a significant step towards addressing the challenges associated with open data quality and maximizing its potential benefits. The increasing reliance on open data for decision-making and the numerous issues surrounding its quality make it crucial to have a reliable tool that assesses and improves the usability of open data.

By providing a web-based application that measures the quality of open data according to six standards, "OpenData Insights" offers valuable assistance to business owners and organizations. It enables them to make informed decisions, verify the quality of open data before making it public, and save time and effort in the process. This application will have a positive global impact on decision-making processes, ultimately benefiting companies and government agencies alike.

Moreover, the local impact of the development of "OpenData Insights" aligns with the goals of Vision 2030 in Saudi Arabia. By maximizing the economic impact resulting from open data and enhancing the Kingdom's utilization of open data, this tool contributes to the realization of Vision 2030's objectives. The recent release of the "نضيء" tool by The Saudi Authority for Data and AI (SDAIA) in November 2023 further emphasizes the significance of open data usage and quality improvement in the Kingdom. "نضيء" incorporates mechanisms to ensure the protection of individuals' privacy within the datasets as it requires the users to log in from Nafad (نفاذ) which is a reliable system that contributes to increasing transparency in government operations and building trust between the government and citizens., underscoring its commitment to maintaining data integrity and upholding ethical standards in data usage. However, Noodhia (نضيء) doesn't measure the quality of open data neither committing to the international standards Therefore, our tool complements the objectives of "OpenData Insights", such as the aim to enhance the Kingdom's utilization of open data and drive economic growth in the years to come.

6.2 Problems and challenges encountered during software development.

While creating our website, "OpenData Insights," we faced the challenge of finding a universal formula to measure data quality across different fields. To overcome this, we conducted extensive research, engaged with specialists, and collaborated with relevant individuals. Our goal was to develop a system that could effectively measure the six standards of data quality in diverse open data domains. Despite the obstacles, we managed to simplify the process and consistently measure data quality across all domains. Additionally, during the software development of the project, we encountered various problems and challenges that required innovative solutions to ensure the successful completion of "OpenData Insights", and those include:

1. Topic's novelty to the Development Team: The subject matter of the project is new to the development team, and it is considered a novelty in the field.
2. Limited Availability of Resources: There are few available resources on the topic, making it challenging to access relevant information and references.
3. Limited time to implement the project: There wasn't much time to learn the Flask framework, Python and its associated libraries like Pandas, The Elementtree XML, and openpyxl. It took a great deal of coordination to ensure both the satisfaction of the client and the efficiency of the work considering the strict timeline of deliverables in an academic context.
4. Delayed Implementation Start: The team faced difficulties in initiating the implementation process promptly due to the ambiguity surrounding certain standards and the scarcity of available resources.
5. Lack of National Methodologies in Measuring Data Quality: As of the date of this report, there are no published national standards for open data sets. Consequently, we rely on international standards, which lack clear definitions.
6. Managing Dynamic Website Link Changes: The challenge arose from the dynamic nature of the Saudi open data portal's website link. As we developed features reliant on this link, such as reliability and timeliness in data retrieval, its frequent changes caused system crashes. Despite our coding efforts, the link's instability posed a significant obstacle.

6.3 Limitations of the system.

Currently, our system has limitations in terms of the file formats it supports. Specifically, it can only handle CSV, XML, JSON, and Excel files. This means that users need to convert their datasets into one of these supported formats before they can be uploaded for data quality measurement. Unfortunately, this restriction hinders the system's ability to analyze data from diverse sources, as it cannot directly process datasets in unsupported file formats such as PDF and HTML.

6.4 The main contribution of the project

The main contribution of this project is the development of "OpenData Insights," a powerful tool that assesses the quality of open data using six open data standards and provides recommendations on suitable standards that measure the quality of the uploaded open dataset. By providing a free, reliable, and efficient way to evaluate open datasets, this project enhances usability, supports decision-making, and drives economic growth, ultimately benefiting businesses, organizations, and the broader society.

6.5 Future work.

One of the future goals is to give the user the chance to define the optimal quality and set up the threshold. This can be done via several ways, such as using the Generative AI to generate the target quality level based on the extracted domain from the given data. Additionally, we aim to offer users the option to share and discuss their results within the system. Presently, users can share results by downloading a PDF file and attaching it to an email.

Our future plans involve expanding the scope, such as enhancing the efficiency of the website by enabling the system to accept a wider range of file extensions, accommodating a broader range of user needs.

7. Acknowledgments

Firstly, we thank God for everything, we couldn't have achieved what we have achieved without the help of Allah.

Secondly, we would like to extend our sincere thanks to Dr. Luluh Aldhubayi for her valuable guidance, great supervision, and support during the entire project. Her experience and guidance have been instrumental in the success of OpenData Insights.

Moreover, we are grateful to KSU Data Center and SDAIA for their kind cooperation, valuable meetings, and sharing their expertise in our project's field. Their support has enabled us to gain crucial insights and successfully complete our system.

Additionally, we express our appreciation to King Saud University, especially the Information Technology department in the College of Computer and Information Science. The knowledge and skills we acquired during our studies in them played a significant role in the successful completion of this endeavor.

Last but not least, we thank our families and friends for believing in our abilities, supporting us, and encouraging us throughout the project journey.

{وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ}

تم بحمد الله! والله ولي التوفيق.

8. References

[1] S. D. a. A. I. Authority, "Open data," [Online]. Available: https://sdaia.gov.sa/en/SDAIA/eParticipation/Pages/OpenData.aspx. [Accessed 28 august 2023].

[2] SDAIA, "# سدايا وبرنامج 'التحول الوطني' يطلقان أول مؤشر وطني للبيانات، ومنصتي (البيانات المفتوحة) و (حوكمة البيانات) #المنتدى_السعودي_للبيانات غدًا الاثنين.," Twitter, 26 November 2023. [Online]. Available: https://twitter.com/sdaia_sa/status/1728699340888093149?s=48&t=9k4PZgzOfEuaiur7Zsz4HQ. [Accessed 30 November 2023].

[3] بوابة البيانات المفتوحة | "عن بوابة البيانات المفتوحة, ا. المفتوحة,ب. ا," [Online]. Available: https://od.data.gov.sa/ar/about/open-data-portal. [Accessed 16 September 2023].

[4] "بوابة البيانات المفتوحة, ا. ا. للبيانات,ا," [Online]. Available: https://data.gov.sa/. [Accessed 16 September 2023].

[5] "Data Quality Guideline: Open data portal," Data Quality Guideline | Open Data Portal, [Online]. Available: https://od.data.gov.sa/en/guidelines/data-quality. [Accessed 1 December 2023].

[6] A. C. L. T. M. M. D. I. R. &. M. F. Vetrò, "Open Data Quality Measurement Framework: Definition and Application to Open Government Data.," February 2016. [Online]. Available: https://www.researchgate.net/publication/295394863_Open_data_quality_measurement_framework_Definition_and_application_to_Open_Government_Data?enrichId=rgreq-c439c292bcf1c92cb1869361ae74e5f3-XXX&enrichSource=Y292ZXJQYWdlOzI5NTM5NDg2MztBUzozNDQ4OTMyNjUzMzQy. [Accessed 21 September 2023].

[7] "CcNSO members," Country Code Names Supporting Organisation, [Online]. Available: https://ccnso.icann.org/en/about/members.html. [Accessed 2 December 2023].

[8] R. Python, "Python's Requests Library (Guide),," Real Python, 2022. [Online]. Available: https://realpython.com/python-requests/. [Accessed 1 December 2023].

[9] "Tldextract," PyPI, [Online]. Available: https://pypi.org/project/tldextract/. [Accessed 1 December 2023].

[10] B. In, "What Is Web Development? (Definition, Types, Career) | Built In," [Online]. Available: https://builtin.com/software-engineering-perspectives/web-development. [Accessed 20 September 2023].

[11] *What is Flask Python*. (2021). Retrieved from Python Tutorial: https://pythonbasics.org/what-is-flask-python/

[12] "Apache ECharts," [Online]. Available: https://echarts.apache.org/en/index.html. [Accessed 3 November 2023].

[13] E. D. Experts, "What is The ChatGPT API: An Essential Guide," [Online]. Available: https://blog.enterprisedna.co/chatgpt-api/. [Accessed 16 September 2023].

[14] M. Sakpal, "12 actions to improve your data quality," Gartner, [Online]. Available: https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality. [Accessed 20 September 2023].

[15] C. Roberts, "5 reasons why data accuracy matters for your business," Medium, 26 October 2019. [Online]. Available: https://chrisrob978.medium.com/5-reasons-why-data-accuracy-matters-for-your-business-b490d5e20bf1. [Accessed 20 September 2023].

[16] R. Fernandez, "A comprehensive guide: How to measure data quality," TechRepublic, 3 October 2022. [Online]. Available: https://www.techrepublic.com/article/how-to-measure-data-quality/#tools. [Accessed 20 September 2023].

[17] "Analytic apps with dataiku," Dataiku, 2023. [Online]. Available: https://www.dataiku.com/product/key-capabilities/analytic-apps/. [Accessed 20 September 2023].

[18] "Ataccama Data Quality," Ataccama, [Online]. Available: https://www.ataccama.com/platform/data-quality. [Accessed 20 September 2023].

[19] "Talend Open Studio: Open-source ETL and free data integration Talend," [Online]. Available: https://www.talend.com/products/talend-open-studio/. [Accessed 20 September 2023].

[20] F. Filip, "Comparison of free and paid versions of the Talend Platform," MIM, 2023. [Online]. Available: https://www.mim.sk/index.php/en/news-en/item/247-comparison-of-free-and-paid-versions-of-the-talend-platform. [Accessed 20 September 2023].

[21] "Check Out This Dataiku Demo," Dataiku, [Online]. Available: https://pages.dataiku.com/experience-a-dataiku-demo?utm_id=14648494444--127992690635--545837670716--dataiku&utm_source=emea-adwords&utm_medium=paid-search&utm_campaign=GLO+Product+Demo&gad=1&gclid=EAIaIQobChMIsaKSyfGzgQMV8YpoCR1tuwP0EAAYASABEgLJYvD_BwE.. [Accessed 20 September 2023].

[22] "Data Quality - Discover Dataiku," Dataiku, 2023. [Online]. Available: https://discover.dataiku.com/data-quality/. [Accessed 20 September 2023].

[23] I. Gartner, "Dataiku review in Multipersona Data Science and Machine Learning Platforms,," Gartner, 2023. [Online]. Available: https://www.gartner.com/reviews/market/multipersona-data-science-

and-machine-learning-platforms/vendor/dataiku/product/dataiku/review/view/4912470.. [Accessed 20 September 2023].

[24] Informatica, "Cloud data quality - Data Quality Management Tool," Informatica," Informatica, 2023. [Online]. Available: https://www.informatica.com/products/data-quality/cloud-data-quality-radar.html. [Accessed 20 September 2023].

[25] *What is Scrum?* (2024). Retrieved from ScrumAlliance: https://www.scrumalliance.org/about-scrum

[26] *Three Pillars of Scrum: Understanding Scrum's Core Principles*. (2024). Retrieved from Attlasian: https://www.atlassian.com/agile/project-management/3-pillars-scrum

[27] Fiix, "System Availability: How to Master Availability and Other Maintenance Metrics," Fiix Software, 2023. [Online]. Available: https://fiixsoftware.com/glossary/system-availability/. [Accessed 1 December 2023].

[28] J. Nielsen, "Do Interface Standards Stifle Design Creativity?," in Nielsen Norman Group, [Online]. Available: https://www.nngroup.com/articles/do-interface-standards-stifle-design-creativity/. [Accessed 19 November 2023].

[29] R. Ramandeep, "Python UUID Generator Program," Scaler Topics, [Online]. Available: https://www.scaler.com/topics/uuid-generator-python/. [Accessed 2 December 2023].

[30] "Pandas documentation - pandas 2.1.3 documentation," [Online]. Available: https://pandas.pydata.org/docs/. [Accessed 13 November 2023].

[31] OpenPyXL, "A Python library to read/write Excel 2010 xlsx/xlsm files," 2023. [Online]. Available: https://openpyxl.readthedocs.io/en/stable/ . [Accessed 30 November 2023].

[32] "xml.etree.ElementTree - The ElementTree XML API Python documentation," Python Documentation, [Online]. Available: https://docs.python.org/3/library/xml.etree.elementtree.html. [Accessed 13 November 2023].

[33] I. M. (Makaranka), "Guide to data quality management: Metrics, process and best practices," Guide to Data Quality Management: Metrics, Process and Best Practices," [Online]. Available: https://www.scnsoft.com/blog/guide-to-data-quality-management. [Accessed 13 September 2023].

[34] [. H. H. a. E. S. Silva, "The role of chatgpt in Data science: How ai-assisted conversational interfaces are revolutionizing the field," MDPI, [Online]. Available: https://www.mdpi.com/2504-2289/7/2/62. [Accessed 13 September 2023].

[35] "Reading an Excel file using Python," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/reading-excel-file-using-python/. [Accessed 30 November 2023].

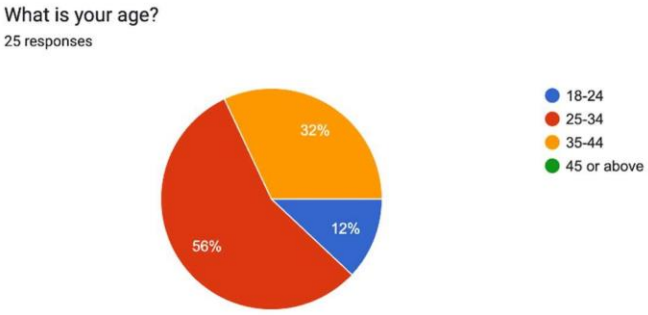
[36] "Draw.io - free flowchart maker and diagrams online Flowchart Maker & Online Diagram Software," [Online]. Available: https://app.diagrams.net/ . [Accessed 1 December 2023].

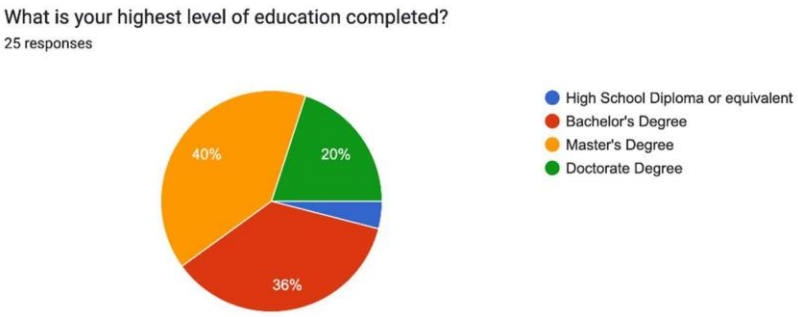
[37] K. S. Rubin, Essential Scrum: A practical guide to the most popular agile process, Upper Saddle River: NJ: Addison-Wesley, 2017.

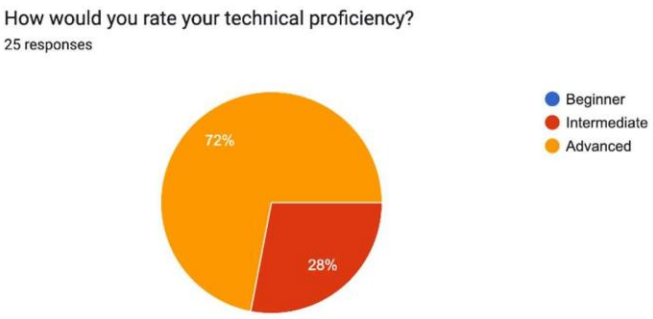
9. Appendix

9.1 Appendix A

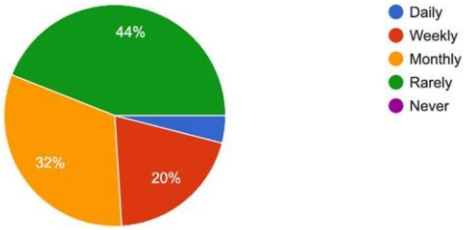
Experts' questionnaire



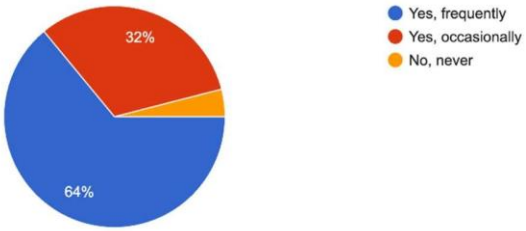




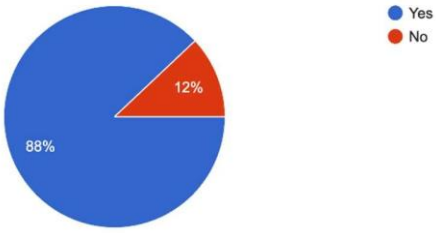
How often do you rely on open datasets for your work or projects?
25 responses



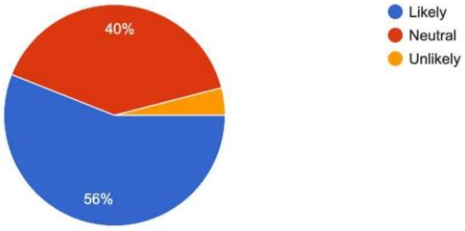
Have you ever encountered issues or challenges due to poor quality in open datasets?
25 responses



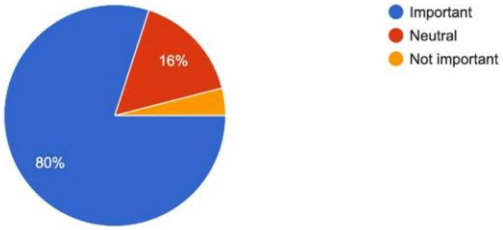
Does measuring dataset quality take a lot of time?
25 responses



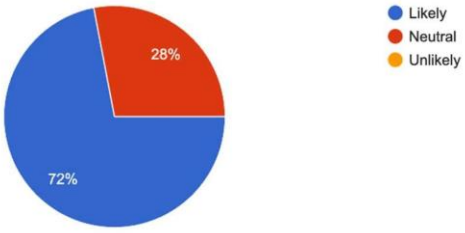
How likely are you to recommend high-quality open dataset results with others?
25 responses



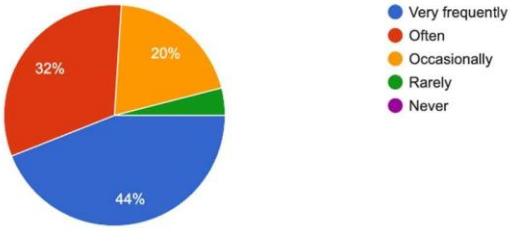
How important is it for open datasets to be sourced from reputable and trusted organizations?
25 responses



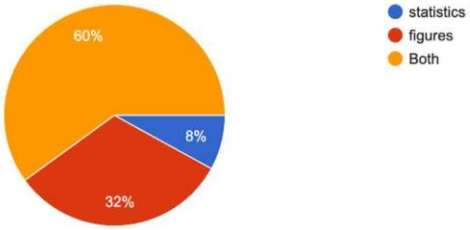
How likely are you to switch to alternative open datasets if you encounter quality issues with your current dataset?
25 responses



How frequently do you rely on visualization tools or dashboards to explore and analyze open datasets?
25 responses



Do you prefer statistics or figures?
25 responses



9.2 Appendix B

Experts' Interviews and Answers	
Question 1: What is your Name, age, job title, level of education, and technical level?	• Bayan Ibrahim 33 years old, master's degree, Data Visualization Department Manager and lead data scientist, Advanced.
Question 2: When you hear the term "open data" what comes to mind? How would you describe it in your own words?	• Accessible and useful data
Question 3: When you think about a good open dataset, what information or details would you expect to find in it?	• Metadata; i.e. the meaning of the columns, any acronyms, and assumptions. It is also helpful to know the date it was collected and the source.
Question 4: Have you ever used open data before? If so, could you share an example of how it was helpful to you or others?	• Yes, I have.
Question 5: Do you use any tools for measuring the quality of data?	• Yes, I use libraries that give me descriptive statistics on the data that produce certain metrics to help me quickly identify the size, distribution, and trend in the data, as well as identify the quality issues. I also use data visualization tools to help me quickly

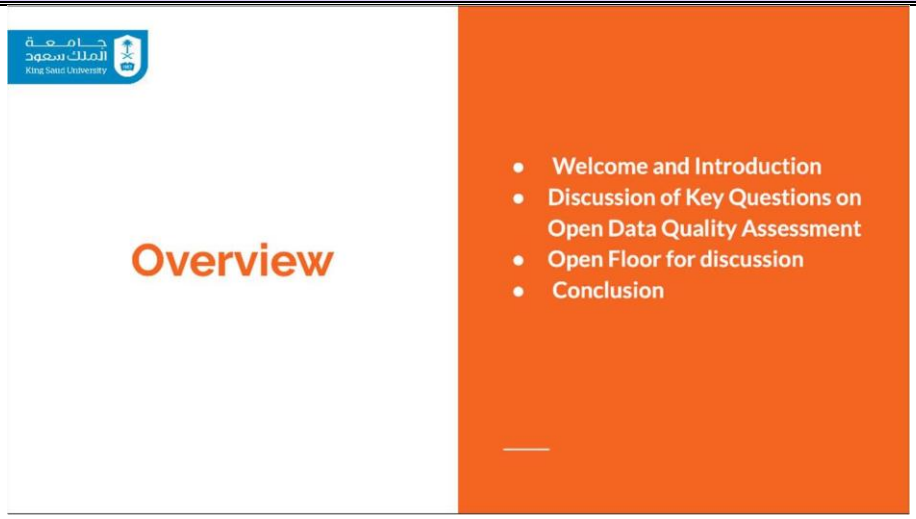
	explore and understand the data like MatPlotLib ,Plotly and Seaborn.
Question 6: Do you find it challenging to ensure the quality of the open datasets?	• Yes, I often spend a lot of time in the process of data wrangling/cleansing, and sometimes I'd have to do go through the process of scraping data online to complement the shortcomings of the open dataset that I've found, or decide the data is not fit for the purpose I have and not use it.
Question 7: Does it take a long time to make sure that your open data is of high quality?	• Yes, and it's almost never is of high quality, but I try to reach an accepted quality where it is fit for the purpose I am using it for.
Question 8: When evaluating the timeliness of open data, what factors would you use to determine its timeliness? How would you handle situations where data is not updated or published in a timely manner?	• I would try to understand the trend in the available dataset and use statistical methods to predict the value during unknown periods. If my question isn't affected by and the observation I'm trying to analyse doesn't usually change much, limited time periods would still be acceptable.
Question 9: What features or functionalities would you like to see in a dashboard that assesses the quality of open datasets? How would you prefer the information to be presented?	• I'd like to know the following; 1. For each column, the percentage of missing data shown as a bar. 2. For each column, the most occurring and least occurring value and the occurrence percentage of each.

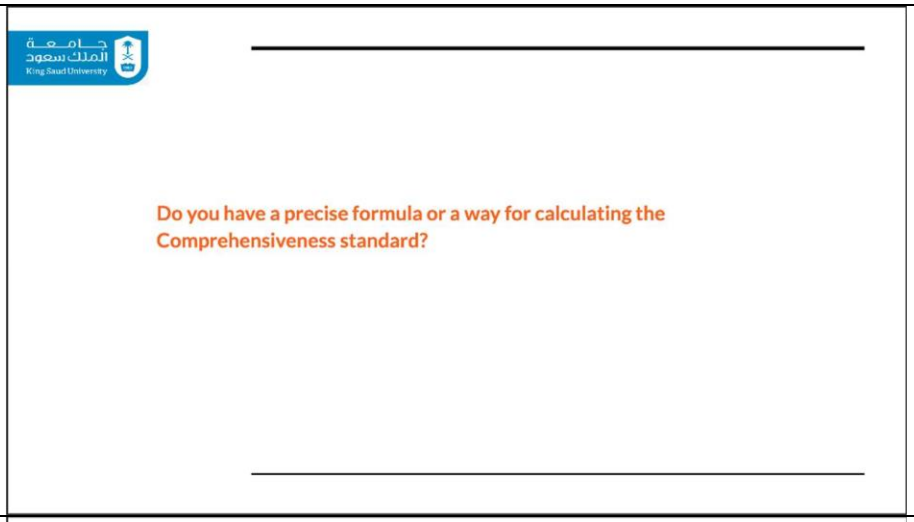
	3. For each column, understand the values distribution by seeing a histogram. 4. The number of rows that are duplicates and their percentage from the overall dataset. 5. The anomalies or outliers.
Question 10: What non-functional features do you prefer the most or you think are the most important?	• A good level of assessment accuracy, and reasonable time to process the dataset and be presented with results/recommendations, e.g, not more than 3-5 minutes.

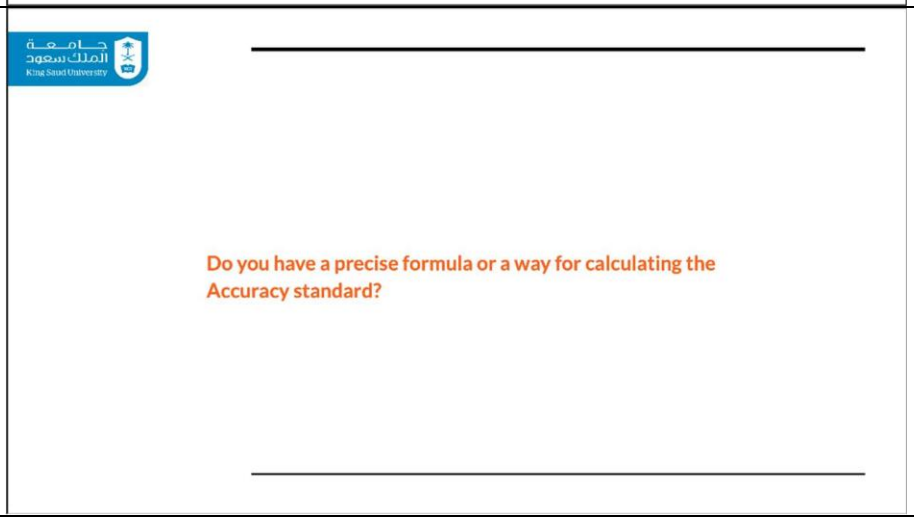
Experts' Interviews and Answers	
Question 1: What is your Name, age, job title, level of education, and technical level?	Saif Alsaif, 35 years old, PhD holder, Artificial Intelligence Consultant, Expert.
Question 2: When you hear the term "open data" what comes to mind? How would you describe it in your own words?	Comes into mind a data set that is open to the public and can be used to get insights.
Question 3: When you think about a good open dataset, what information or details would you expect to find in it?	A good open data set is a clean data set. It must be comprehensive. It is easy to use and to download, transfer, and analyze.
Question 4: Have you ever used open data before? If so, could you share an example of how it was helpful to you or others?	I have used tabular data and I have also used visual data. The last data set that I used was a data set on weather forecasting.

Question 5: Do you use any tools for measuring the quality of data?	No.
Question 6: Do you find it challenging to ensure the quality of the open datasets?	Yes.
Question 7: Does it take a long time to make sure that your open data is of high quality?	Yes, it does take a long time and sometimes a small misalignment in the data set could ruin the whole dataset.
Question 8: When evaluating the timeliness of open data, what factors would you use to determine its timeliness? How would you handle situations where data is not updated or published in a timely manner?	Being an engineer, I would use historical data to generate new random data. This will help me deal with the timeliness of data.
Question 9: What features or functionalities would you like to see in a dashboard that assesses the quality of open datasets? How would you prefer the information to be presented?	I would like the dashboard to have a final score of the quality of each factor of the data set, and also to have some visualization of a sample of the data.
Question 10: What non-functional features do you prefer the most or you think are the most important?	The system should be easy for me to understand without wasting my time. Additionally, I want my actions to be implemented quickly, not more than 5 minutes.

Expert Consultation meeting's Agenda







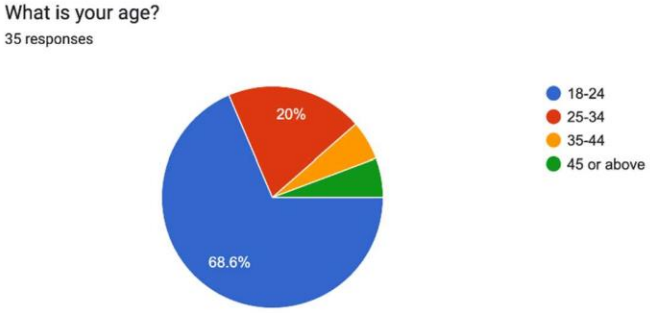
Do you have a precise formula or a way for calculating the Reliability standard?

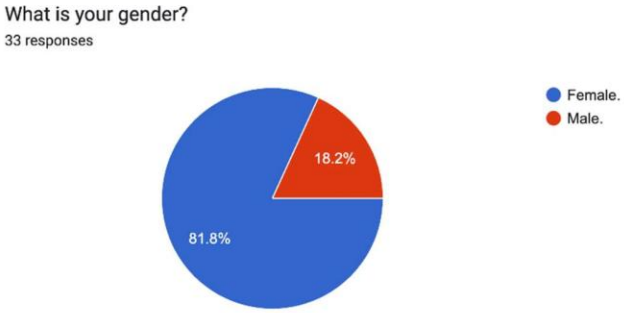
Do you have a precise formula or a way for calculating the Timeliness standard?

What are the cutoff values used to measure the quality of (accuracy, completeness, timeliness, consistency, reliability, and comprehensiveness) standards?

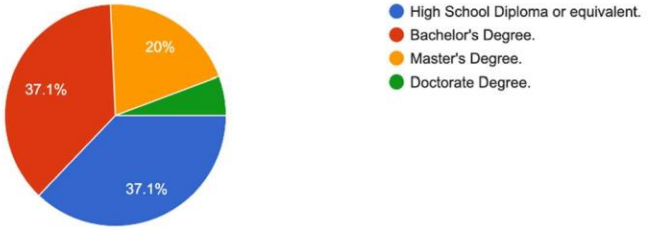
9.3 Appendix C.

Data consumer's questionnaire

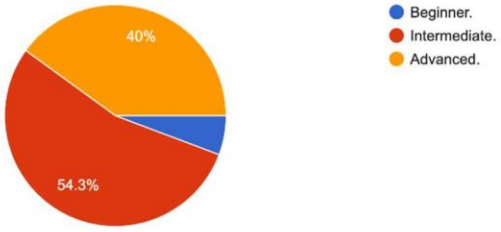




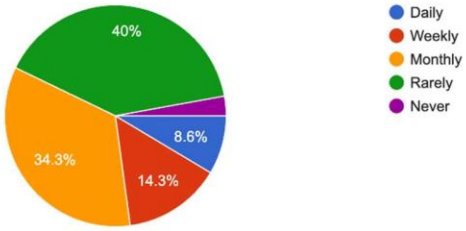
What is your highest level of education completed?
35 responses



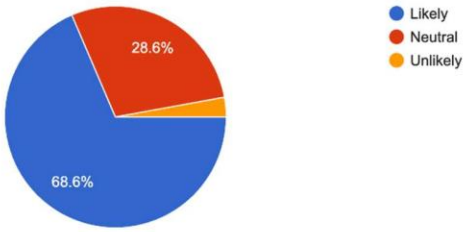
How would you rate your technical proficiency?
35 responses



How often do you rely on open datasets for your work or projects?
35 responses

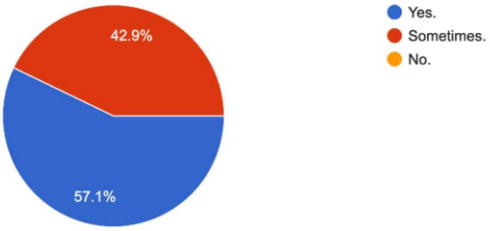


How likely are you to recommend high-quality open dataset results to others?
35 responses



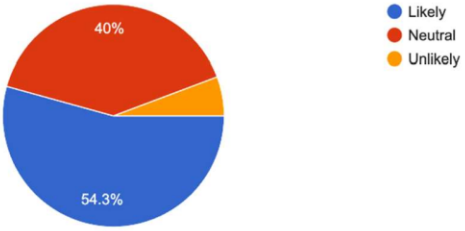
Do you find it difficult to ensure that your open dataset is reliable and of good quality before using it for example to make some decisions or research?

35 responses



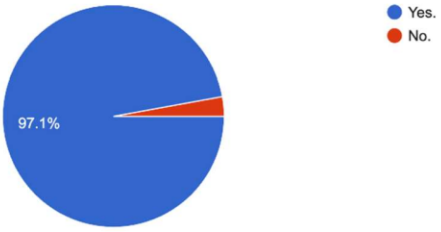
How likely are you to switch to alternative open datasets if you encounter quality issues with your current dataset?

35 responses



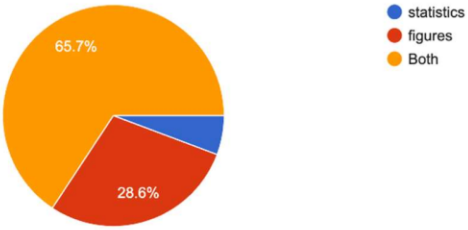
Would you like for the system to recommend similar datasets to the one you upload?

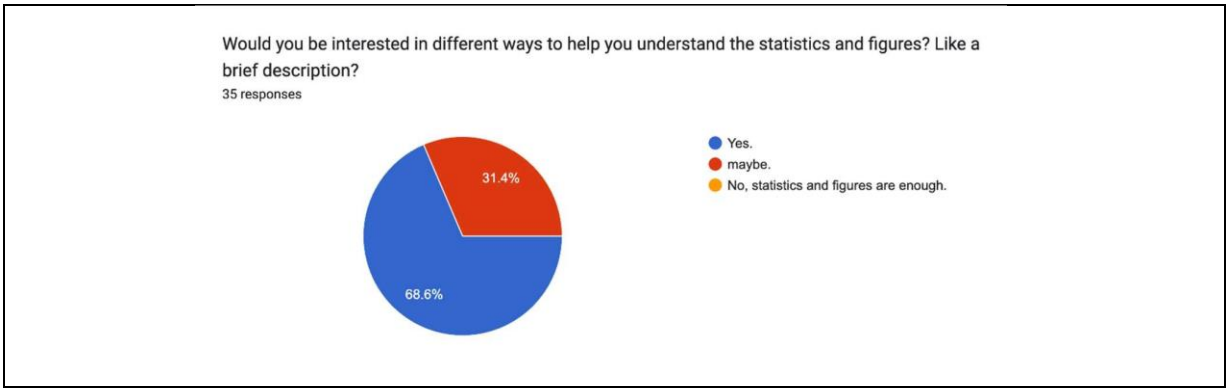
35 responses



Do you prefer statistics or figures?

35 responses





9.4 Appendix E.

Data consumer's Interviews and Answers	
Question 1: What is your Name, age, job title, level of education, and technical level?	• Fai, 23 years old, Software engineer, bachelor's degree, mid-level. • Norah, 26 years old, solutions architect, master's degree, Mid- level.
Question 2: When you hear the term "open data" what comes to mind? How would you describe it in your own words?	• A publicly available dataset that anyone can access and use. • It's the datasets that can be accessed by anyone, from anywhere without any restrictions on their usage.
Question 3: When you think about a good open dataset, what information or details would you expect to find in it?	• I expect it to be organized into categories and everything to be clear, and the data should be complete as well. • That it would be easily understandable, and everything is consistent which means that there aren't any conflicts in the dataset.

Question 4: Do you use any tools to measure the quality of data? If no Would you prefer to use a tool?	• I should do so, but no, I just check it manually, yes, I would love to. • I only view the open datasets, without paying much attention to the quality, because I'm mostly depending on the fact that it's already in a good state once it's published to the public, but I would love to make sure and use a tool.
Question 5: Do you find it difficult to ensure that your open dataset is reliable and of good quality before using it, for example, to make some decisions or research?	• Yes, definitely, and this may affect my work. • I mostly depend on the datasets that are approved by SDAIA, and I don't always find what I'm looking for.
Question 6: How would you prefer the information to be presented?	• I can understand visuals more quickly, so I prefer graphs and charts. • I prefer to use figures only; I don't like to see paragraphs explaining the charts.
Question 7: What features or functionalities would you like to see in a dashboard that assesses the quality of open datasets?	• Since I am looking at charts, that means I am going to make a decision, so I want to be able to make a comparison between multiple datasets I might upload. • I would like an explore button so that I could share the results with others using extensions such as csv and excel files, or even save it in my device if it's a json file. Also, I would like some

	history feature to keep my records in it for future reference.

9.5 Appendix F.

GitHub: https://github.com/OpenDataInsight/2023-GP1-7-Final-Release-1.git

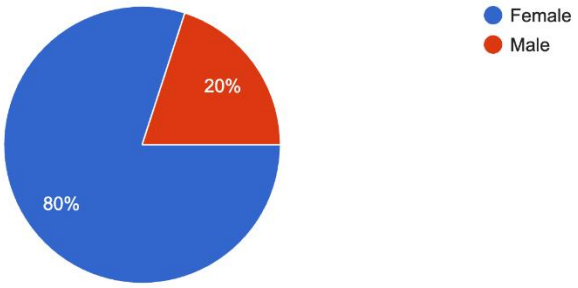
Jira: https://opendata-monitor.atlassian.net/jira/software/projects/GP/boards/1/backlog

9.6 Appendix G.

End user's questionnaire

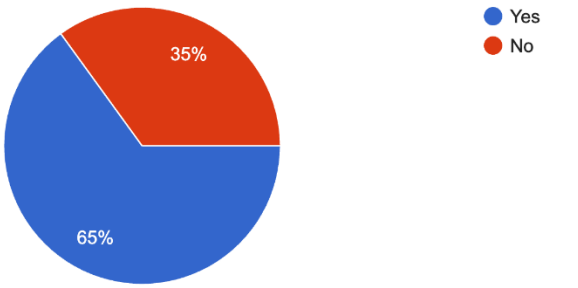
What is your Gender?

20 responses



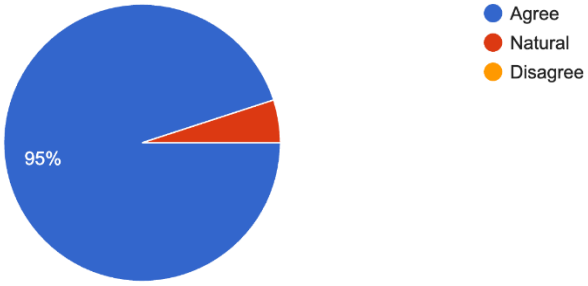
Do you have a technical background?

20 responses



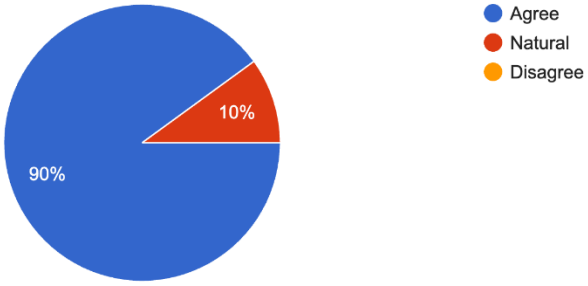
The dashboard's user interface is visually appealing and user-friendly.

20 responses



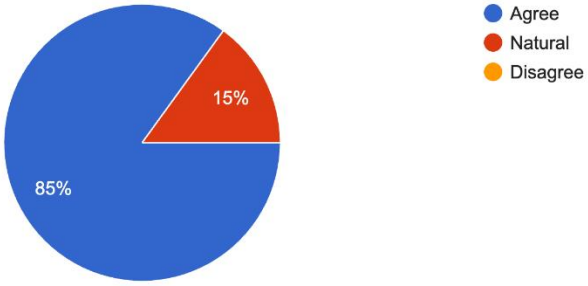
The charts(results) are effective and easy to understand.

20 responses



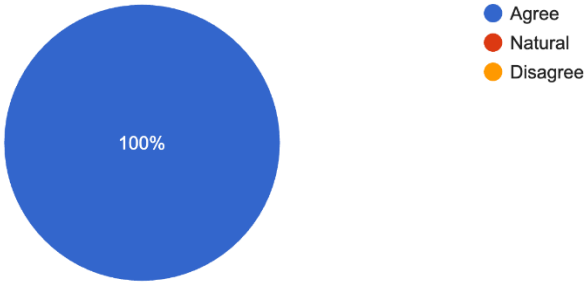
The system is easy to use and navigate.

20 responses



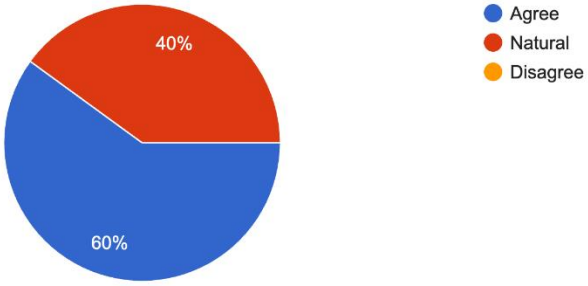
You were able to interpret and understand the completeness result.

20 responses



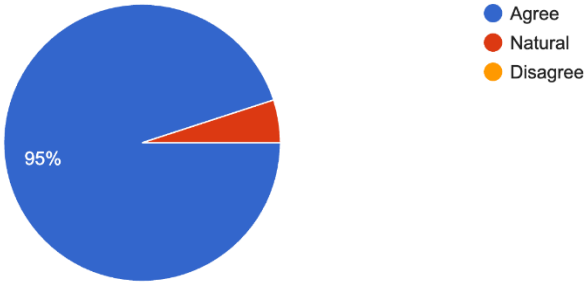
You were able to interpret and understand the Consistency result.

20 responses



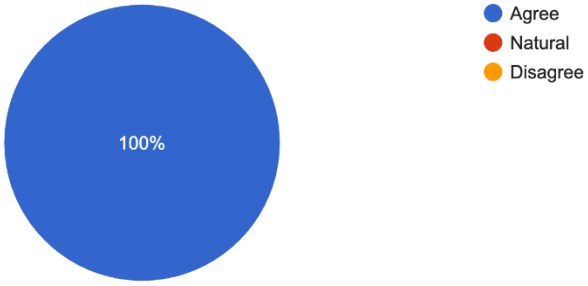
You were able to understand the report.

20 responses



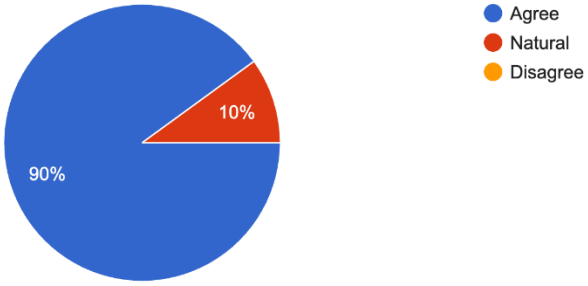
The view profile page is clear and easy to understand and Interact with.

20 responses



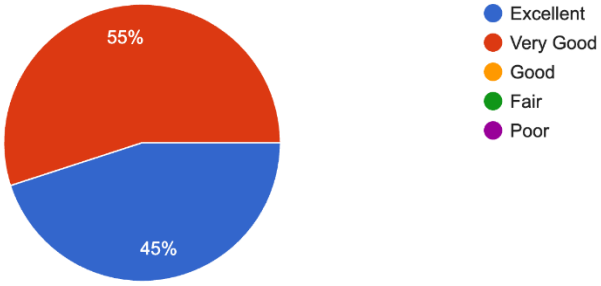
You find the process of checking the reliability of the dataset intuitive and easy to understand

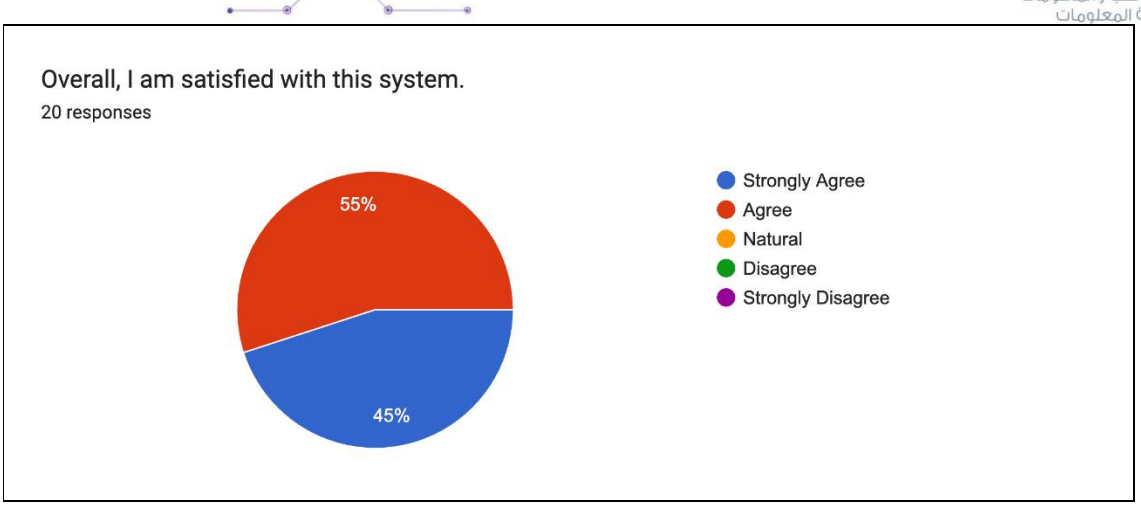
20 responses



How would you rate the overall user experience of the dataset upload function?

20 responses




Overall, I am satisfied with this system.
20 responses
55%
45%
Strongly Agree
Agree
Natural
Disagree
Strongly Disagree