# A comprehensive review of open data platforms, prevalent technologies, and functionalities

Mohsan Ali
University of Aegean, Samos, Greece
Mohsan@aegean.gr

Charalampos Alexopoulos
University of Aegean, Samos, Greece
Alexop@aegean.gr

Yannis Charalabidis
University of Aegean, Samos, Greece
Yannisx@aegean.gr

## ABSTRACT

Open data can play a crucial role in different sectors of the world, such as government, science, research, technology, culture, and finance. There are several necessary measures that every organization needs to consider before opening data. There are three major steps to opening the data: (1) Preparation stage, and (2) launching the open data initiative (3) In this case, the feedback mechanism study such as expand and sustain stage, our focus is on the second step, which is how to launch the data as an open dataset and what platforms are available in the market. There are several pros and cons associated with each platform that need to be discussed before publishing the data to get the maximum outcomes in a sustainable and transparent way. We will discuss seven major open data platforms, such as (1) CKAN (2) DKAN (3) Socrata (4) OpenDataSoft (5) GitHub (6) Google datasets (7) Kaggle. We will evaluate the technological commons, techniques, features, methods, and visualization offered by each tool. In addition, why are these platforms important to users such as providers, curators, and end-users? And what are the key options available on these platforms to publish open data? At the end of this study, an individual will be able to select one of them for their open data initiative to launch the data as open.

## CCS CONCEPTS

• **Applied computing** → Computers in other domains; Computing in government; E-government; Computers in other domains; Digital libraries and archives.

## KEYWORDS

Open data, Catalogs, OGD, OD Platforms, Open data portals, Open data ecosystem

**ACM Reference Format:**
Mohsan Ali, Charalampos Alexopoulos, and Yannis Charalabidis. 2022. A comprehensive review of open data platforms, prevalent technologies, and functionalities. In *15th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2022), October 04–07, 2022, Guimarães, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3560107.3560142


*ICEGOV 2022, October 04–07, 2022, Guimarães, Portugal*

ACM ISBN 978-1-4503-9635-6/22/10…$15.00
https://doi.org/10.1145/3560107.3560142

## 1 INTRODUCTION TO OPEN DATA AND PLATFORMS

Open data is the data which is freely available to use and distribute even though for commercial use. There are several other detailed definitions for open data [1] [2]. One of the most appropriate definitions is developed based on several existing considerations. Open data needs to follow up the steps for it to be completely open [3], [4]. Open data should be complete and must be uploaded in a complete manner. It should be timely for any decision-making. Open data should be a primary focus. It should be collected by experts or researchers. Open data accessibility also matters through open data platforms and catalogs through API's or other 5-star formats. Machine readability is an important aspect to consider, although when data is available in csv or excel, it is considered machine readable. The data must be non-discriminatory without considering gender or any other factor while opening. Non-proprietary means it must be available under open licensing, which means it can be used and distributed by anyone. Free means that data must be available under the two most used licenses, such as CC-BY and CCO, while publishing the open data. The data also needs to be reviewable to remove outliers, errors, and omissions (E&O) when detected by any end user [5] [6].

There are three steps involved in the open data plan[1] which are (1) Preparation of open data for publication (2) launching the open data stage and (3) Extending and sustaining the open data stage [7]. In this research we will just focus on the 1st stage, which will help the publishers to select the proper platform for publishing their data. The preparation stage is a well-defined process such as at this stage we check the open data readiness assessment by evaluating the technology skills, senior leadership involvement, civic engagement, financing for OD, data management plan, data demand, policy and legal perspective of OD, and institutions involvement in OD. After checking the open data readiness assessment, open data providers build the open data inventory where a mapping between supply and demand is drawn. The legal and policy work is also considered at this stage such as open data licenses, open data policies, and legal review. The essential license is selected based on the open data properties and organization rules. The steering committee can be involved to better decide which legal and policy work needed to be more focused. The last stage of preparation of the open data is the establishment of a data catalog to share the open data with users, and community by the means of data curators. Now on word in this study, our focus will be on the open data platforms, portals, or catalogs [8].

There are four types of data catalogs, such as open source, software as a service, federated, and standalone catalogs. There are

[1] https://opendatatoolkit.worldbank.org/en/

specific properties, benefits, pros, and cons associated with each type of data catalog. For instance, open-source data catalogs are flexible and customizable, but on the other hand, strong technical knowledge is required to maintain them[2]. Software as a service catalog are paid versions that are only available after paying a fee to the vendor who manages and maintains the catalog, with fees based on how much open data is used and how much data is needed for that user [9]. The federated catalogs are used to integrate several other organizations' or ministries' catalogs into a single place and can support topic-specific or separate catalogs. The last category of data catalogs is called "standalone," which means that there is only one catalog where anyone can find all the government data [10]. This study focuses on the open-source data catalogs, platforms, and portals that anyone can use without paying a fee and customize according to their requirements. The main objective of this study is to develop an analysis framework for analyzing the different features of data platforms. Furthermore, we proceeded to the usage analysis of these platforms by open data portals around the world. There are several data platforms used to develop open data portals by governments, businesses, and NGOs to share data with the public. After that, Contributions to this research article are:

- Analysis framework to compare the different open data platforms
- To elaborate the open data platforms selection criteria for better data availability, accessibility, and reusability
- This work helps open data publishers get a better idea of how to choose the best open data platforms for publishing open data
- The following sections outline the organization of this paper: The second section discusses the platforms in detail and what has been done so far. The third section demonstrates the features, technologies, and functionalities provided by the CKAN, DKAN, GitHub, OpenDataSoft, Google datasets, Kaggle, and Socrata. The fourth section will explore the critical role of these platforms in ensuring the long-term sustainability and reusability of open data. The final piece is the conclusion, in which we will determine which of these data platforms is the best and why.

## 2 LITERATURE REVIEW

This section will elaborate the open data background and importance of platforms in the field of open data.

### 2.1 Open data background

The openness of the data really matters because openness decides the use, reuse, and redistribution of the dataset in the real world. Tim Berners-Lee's (inventor of the web and linked data initiative) Five Star[3] description for open data can be explored with respect to availability and usability dimensions. In a five-star model of linked data, every star shares property from the previous stars and has their own unique attributes as well. Table 1 illustrates the 5-star linked data properties, explanation, complexity for the publisher and consumer, and type of license. There are many digital licenses available in open data, but we only talked about the most common ones[4]. There are different maturity models used in the research to assess the commitment and capabilities of public, government, and private agencies in the field of open data. During the phase of publishing open data, data publishers also consider things like legal, institutional, technological, and citizen and business perspectives [6]. Lourenço et al. describes another maturity model, to identify relevant contextual components. There are factors that influence how public institutions publish data on their portals. The last few components are then put together into an online transparency for an accountability maturity model, which is used to measure how developed a region is [11].

### 2.2 Data Platforms, portals, catalogs, and metadata schemas

Data platforms are central resources to share the data and they play a very crucial role in sustainable open data efforts such as CKAN and DKAN. Open data portals are based on data platforms and are the central repositories where data can be found, re-used, and redistributed according to licensing. Some platforms convert data from one format to another, such as from CSV to XML, JSON, Excel, etc. [12]. Open data is more focused towards the high availability and openness of governmental data, there should be focus on the open government metadata. Open data portals provide the metadata features according to their own. Several organizations are collaborating to make the metadata standards to resolve the metadata quality issues [5]. EU Digital Agenda[5] states that, Semantic interoperability and metadata quality or management are closely linked. If metadata is not in good quality, then semantic interoperability of open data would be weak and vice versa. The European interoperability framework defines semantic interoperability as "the meaning of the open data elements, s and the relationship between them. It includes developing vocabulary to describe data exchanges and ensures that data elements are understood in the same way by communicating with parties". Metadata schema (MS) can help to reduce the integration, development, and information sharing costs. MS has the capacity to improve interoperability across systems with similar or shared metadata. Interoperability solutions for European public administrations represented five levels of maturity of metadata management [13]. The maturity levels of metadata are important for the analysis of different open data platforms.

### 2.3 Open data portal developments around the world

Data portals are an essential component of open government data initiatives, and the literature provides several solutions with varying characteristics. Government Data Catalogs or Metadata Portals (Repositories) are indexes that include structured descriptions (metadata) of the real data (e.g., data.gov.tw). These techniques have the potential to increase the discoverability of published datasets, as the discoverability of data is strongly related to the quality of the metadata [14]. An open government catalog would include a list of metadata pieces detailing open government datasets as well as online resource links. Although the development of a catalog raises a significant question: "What metadata should be stored and how

[2]https://opendatatoolkit.worldbank.org/en/technology.html
[3]https://5stardata.info/en/
[4]https://joinup.ec.europa.eu/sites/default/files/inline-files/W3C04.pdf
[5]http://ec.europa.eu/information_society/digital-agenda/index_en.htm

should it be presented?" In the context of harvesting, this question is particularly relevant because metadata structure and meaning are not always consistent or self-explanatory [15]. To resolve the problem inconsistent of metadata structure and meaning there are different software used such as CKAN[6], and DCAT[7] vocabulary. DG CONNECT[8], the EU Publications Office[9], and the ISA Program collaborated to develop the DCAT-AP specification. DCAT-AP is widely used by several government portals; one of the examples is Ireland open government data portal (data.gov.ie) [16]. European Union Member States, EU institutions, and the United States worked together to develop the specification. Using the Data Catalog vocabulary (DCAT) developed by the W3C, the DCAT Application profile for data portals in Europe (DCAT-AP) is a standard for characterizing public sector datasets in Europe. Cross-data portal search and enhanced searchability of public sector data across borders and sectors is its primary use case [9]. This can be accomplished through the sharing of data set descriptions across data aggregators. Therefore, the searchability and metadata(specification) of the data in an open government portal or open data portal is a crucial consideration when selecting a portal for data publication [16], [17], [18], [19]. Yang et al. discuss the importance of open dataset categorization for sustainability and value creation in their literature review, but they do not discuss open data portals and their characteristics for publishing open data in the best way [20]. Another interesting article in the open data portals domain is Transparency by Design and Role of open data portals by the Lnenicka et al. and they figured out that what properties or features of the portals leads to the transparency of the open data portals [19]. Our study is completely based on the open data portals and their features, which are different from those dataset repositories which are just used to publish datasets like GitHub, Google Datasets, and Kaggle. Keeping in mind all of the features and characteristics we've talked about in this section, like metadata, categorization, data visualization, etc., we can go into more detail about the features and characteristics of the open data portal in order to evaluate how well they work.

## 2.4 Various open data publishing platforms

CKAN, DKAN, Socrata, OpenDataSoft, GitHub, Google, and Kaggle datasets are just a few examples of open-source catalogs that can be used to share open data with the public [10]. Several open data platforms are available freely to use and develop open data catalogs and initiatives for governmental and non-governmental organizations, businesses, researchers, and NGOs.

*2.4.1 CKAN, DKAN, Socrata and OpenDataSoft.* Comprehensive Knowledge Archive Network (CKAN) is a fully functional, featured, advanced and reliable open-source data management system for sharing and distribution of open data. CKAN provide several solutions such as one can develop national, international, federated platform using the extend functionality of the CKAN[10]. Hundreds of extensions available to make the open data portals more sophisticated and sustainable. CKAN is an open-source DMS that provides data hubs and data portals. CKAN facilitates data publication, sharing, and utilization. For instance, one may utilize the metadata that is available inside CKAN, and if a data provider wants to increase the quality of the metadata, the CKAN+DCAT [11] feature is also available. Furthermore, CKAN provides the following: API's, data store (for the storage of open data), extensions, geospatial (to preview the data), data management tools, search options, visualizations, and professional themes. CKAN can be used for governmental datasets, catalogs, and portals as well as for enterprises such as energy sectors, pharmaceuticals, and finance sectors to upload their assets. CKAN covers two of the largest open data portal projects, such as the UK[12] and US[13] data portals. DKAN is most like CKAN, there are a few distinctions, such as DKAN's usage of the PHP and Drupal. Those organizations whose CMS uses PHP and Drupal can immediately implement the DKAN. Another significant distinction between CKAN and DKAN is the integrated CMS. Those utilizing CKAN, for instance, frequently add another content management system such as Drupal, WordPress, PHP, or Django, whereas DKAN has an integrated content management system. DKAN uses datasets as content to improve information processes and create an open data environment that can last [21]. Most prominent users of DKAN are the U.S. Department of Health & Human Services[14], the U.S. Department of Agriculture[15], Oklahoma's open data[16]. Socrata distributes or maintains more than one hundred open data catalogs from governments, non-profit organizations, and non-governmental organizations around the world. Socrata Open data is publicly accessible and may be redistributed without restriction. Socrata offers a graphical user interface (GUI) for searching and an application programming interface (API) called the SODA API. This makes it possible for apps to be made in the future that give more accurate results. Socrata provides different institution towards the open data ecosystem sustainability they provide "Data and Insights"[17] for the data insight, public engagement, and to optimize the performance [10]. OpenDataSoft (ODS) provides the open data hub for the open data catalogs from multiple areas, such as governments, energy, utilities, banking, insurance, transportation, mobility, public institutions, local governments, services, and manufacturing. The most prominent features of ODS are the open data catalogs from governments, energy, utilities, banking, insurance, transportation, mobility, and public institutions. The most well-known services offered by ODS include the data hub, ODS academy, assistance or support for its customers, code library, training, blogs, books, manuals, and success stories [22]. ODS also uses different interoperability standards, such as DCAT-AP, to make sure that data can talk to each other in a meaningful way. This helps to improve the world of linked open data [23].

*2.4.2 GitHub, Google Datasets, and Kaggle.* Different open data platforms, catalogs, CMS, and data management systems use

[6] http://ckan.org
[7] https://data.gov.ie/dataset/dcat-ap/resource/bd47f44c-0ad7-4002-b601-169a4d1a5e26
[8] https://knowledge4policy.ec.europa.eu/organisation/dg-cnect-dg-communications-networks-content-technology_en
[9] https://op.europa.eu/en/home
[10] https://extensions.ckan.org/
[11] https://extensions.ckan.org/extension/dcat/
[12] https://data.gov.uk/
[13] https://data.gov/open
[14] http://www.healthdata.gov/
[15] https://data.nal.usda.gov/
[16] http://data.ok.gov/
[17] https://www.tylertech.com/products/data-insights

GitHub to make their data and code freely available to the public. Stanford open data is shared via GitHub[18] as well as their own developed portal[19]. There are plenty of solutions which are available on GitHub, but one cannot use GitHub to construct linked open data because it does not support RDF and DCAT-AP based vocabulary, and more than one dataset cannot be linked together based on their relations. GitHub can be used as an open data portal, platform, catalog, or repository management system. Also, GitHub only lets you publish data and is driven by data, so it can't be used as a user-centered open data portal [2]. Using Google Dataset Search, scholars may search the web for publicly available data. The service was introduced on September 5, 2018, and the company stated that it was aimed at data journalists and scientists. The Google datasets are unable to construct links between different datasets based on RDF, vocabulary, or metadata, like CKAN and DKAN. Google Datasets also offer the ability to get data more quickly. Google datasets include csv, xls, xlsx, and many more forms of data. For example, we may use the Google dataset search engine to get datasets from various countries by searching for the data.gov phrase. Different open data portals from different governments could trade links to Google datasets to make the portals more useful and popular [24]. Kaggle is an essential part of any discussion of data repositories. Researchers, academics, and corporations use Kaggle as a data repository and competition manager as well. Kaggle is a great place to learn the fundamentals of data science by finding free datasets, source codes, competitions, and the best competitors. Users get access to over 50,000 public datasets and 400,000 public notebooks. On Kaggle, many developers preferred GUI-based code over API-based code because it provides more information about publications and more options for interacting with data and codes. Kaggle also has an integrated Python IDE, but you can also use other languages for data science, like R, in the Kaggle IDE[20].

## 3 PROPOSED METHODOLOGY AND RESULTS

The proposed methodology for the analysis of seven different open data portals is divided into three parts as shown in Figure 1. In the first part, we extracted different properties to develop the analysis framework for open data platforms, which are used directly or indirectly by different governmental and non-governmental organizations. The analysis framework consists of terms such as metadata, easy access, API, Metadata levels, Searchability, machine readability, etc. the platforms are analyzed based on these 31 properties. These properties are derived very carefully from our best knowledge and available literature in the open data field, such as 5-star linked data by Berners Lee [18], maturity level of metadata [22], and FAIR of open data; technological aspects of open data [18]; user-interfaces; security; standards [2], transparency of OD portals [19], categorization of open data [20], licenses [25], and policies [3]. The second part of this study as per Figure 1 is the identification of open data platforms and documentation of their characteristics. DKAN, CKAN, Socrata, OpenDataSoft (ODS), Google Datasets, Kaggle, and GitHub are identified as open data platforms at large scale. the documentation, literature and authors knowledge utilized to draw the functionalities of each platform based on the properties identified in the previous step. For instance, each tool is evaluated based on maturity of metadata and so on. To check the technological commons and functional differences among these platforms, we apply all the properties one by one. Final step of the proposed methodology, after evaluating the open data platforms based on functionalities and properties, a dataset is extracted from world data portals organization[21]. DataPortals.org is the most exhaustive directory of open data portals in the world. It is curated by a group of famous open data specialists from around the globe, including representatives from local, regional, and national governments; international institutions such as the World Bank, and various non-governmental organizations. At OKCon[22] In 2011 in Berlin, the alpha version of DataPortals.org was introduced. This is an open license portal and is also available on GitHub[23] issue. Total 593 open data portals extracted from national, international, national, and federated to identify their sources to perform analysis among them. The descriptive statistics are applied to the 593 portals to extract important information, such as who are the owners of these data portals? such as how they developed the portals (such as standalone or based on third-party platforms)? What are the licenses that they commonly use? What are the languages that each platform supports? How many portals does each country have in the list of 593 portals? With the help of this strategy, open data providers, users, and policy makers can decide better on the selection of open data publishing, open data use, and decision making respectively.

## 4 RESULTS

Table 1 explains all the features, functions, properties of the open data portals which are important to consider for publisher, users, and policy makers. There are 31 parameters that we extracted by reading the extensive literature about open data portals specification and functionalities. The (+) sign means that a particular portal fulfills the functions mentioned in the first column and (–) sign means platform is not aligned with this functionality or feature. Extra information such as definitions of different terminologies is provided along with the detailed analysis and comments section for each platform. The common functionalities that all platforms provide are metadata, machine-readable, supply and quality of data, data privacy, security protocols for sharing data, technologies used, web services and GUI, data format, and authorization for data collection and sharing. The government, non-governmental, or NGO wants to select an open data portal based on common functionalities.

After knowing the different attributes of the portals, we extracted the dataset from the world data portals organization. DataPortals.org is the most exhaustive directory of open data portals in the world. It is curated by a group of famous open data specialists from around the globe, including representatives from local, regional, and national governments; international institutions such as the World Bank, and various non-governmental organizations. At OKCon in 2011 in Berlin, the alpha version of DataPortals.org

[18] https://github.com/features
[19] https://github.com/StanfordOpenData/open-data-portal
[20] https://www.kaggle.com/
[21] https://dataportals.org/
[22] https://okcon.org/
[23] https://github.com/okfn/dataportals.org/issues/new

**Table 1: Proposed evaluation matrix for the open data platforms**

| Features/ Properties | Definitions | Detailed Analysis and comments | CKAN | DKAN | GitHub | Google Datasets | ODS | Socrata | Kaggle |
|---|---|---|---|---|---|---|---|---|---|
| Metadata | Open data metadata provides structured information about datasets such as dataset titles, descriptions, created dates, and much more. | Metadata data standards such as DCAT-AP. The metadata quality is an important feature for the open data portal management [9]. | + | + | + | + | + | + | + |
| Easy access | Easy to use, sometimes very difficult to access the data because of complex portal user interfaces and filtration of data based on some criteria [26]. | Easy access to data is important for the OD platforms. For those who are unfamiliar with the specific terminologies and functionalities, accessing data on GitHub can be extremely difficult at times. | + | + | - | + | + | + | + |
| Metadata levels | According to ISA metadata levels [13] | GitHub and Google Dataset ISA metadata up to second level. CKAN, DKAN, Socrata, and ODS provides up to $4^{th}$ level of ISA metadata. | + | + | - | - | + | + | - |
| Searchability | Several researchers identify the open data searchability as a barrier and challenges for the end-user [22] such as search by location, by data title, description, or topic wise search | Searchability of dataset is important. CKAN, DKAN, ODS, and Socrata are provides several types of filters and options to search the datasets. GitHub did not have any interface for the data searching. | + | + | - | + | + | + | + |
| API | APIs are essential for generating new products that meet societal needs. To construct a weather forecasting software using open data, one needs understand API availability. Some open data portals offer SPARQL, Restful API, while some don't. [26]. | Google Datasets provide paid version of API to access the datasets. GitHub did not have any specific restful API for the end-users. CKAN, DKAN, ODS, and Socrata provides several types of API's such as restful and custom APIs. Kaggle provides the API based datasets and competition based access. | + | + | - | + | + | + | + |
| Machine-readable data | The 5-Star Berners-Lee linked open data model because it explains how machines can read open data and other related ideas. | GitHub, Kaggle and Google Datasets did not provide the 5-star linked datasets. CKAN, DKAN, Socrata, and ODS are quietly famous for Berners-Lee 5-star linked data. So, before choosing an OD platform, you need to know if it supports linked data. | + | + | - | - | + | + | - |
| Open data Standard compliance | Standards are imposed by different open government organizations and European data standards as well. Public repository and we can apply standards based on organizational rules of data sharing. | GitHub, Kaggle and Google Datasets provide less support for open data standards. Several organizations just share their data with the public without considering the platforms' support for open data standards compliance. | + | + | - | - | + | + | - |

**Table 1: (*Continued*)**

| Features/ Properties | Definitions | Detailed Analysis and comments | CKAN | DKAN | GitHub | Google Datasets | ODS | Socrata | Kaggle |
|---|---|---|---|---|---|---|---|---|---|
| Data visualization | Data visualization is important to get fast institutions from the data. Some portals have this functionality and others just show the metadata information. | Google datasets and GitHub do not provide data visualization. They just provide the interface to upload and view the data. No charts or graphs are provided to get the initial information from the data. CKAN, DKAN, OpenDataSoft, and Socrata provide these functionalities because they also have integrated or extension-based CMS systems. Kaggle provides the data visualization to a limited extent. | + | + | - | - | + | + | + |
| Licenses | Licenses are very important when publishing the data on the open data portals. Mostly used licenses in the open data field are CC0, CC-BY, and open data license of UK [27]. | Licenses are very important in the publishing of open data, data portals, and catalogs. CKAN, DKAN, Socrata, and ODS have almost all types of open data licenses which are non-proprietary. GitHub and Google, and Kaggle Datasets have other types of licenses which might not be open. | + | + | - | - | + | + | - |
| User support | User support is a very useful feature of portals, if efficiently managed. Just a few portals provide this opportunity to maintain the quality of service. In government open data portals user support is very important at large scale. | GitHub, Kaggle and Google datasets are not developed for open data publishing, but several organizations do this. That is why the organizations were not able to provide the users with proper support. Other OD platforms such as CKAN, DKAN, Socrata, and ODS have the proper user support. | + | + | - | - | + | + | - |
| Open data categorization | Open data categorization is a valuable property/functionality of the open data platforms where different datasets are categorized into different distinguished types, e.g., agriculture, finance, and energy. | On one hand, DKAN, CKAN, Socrata, and ODS have this functionality, but on the other hand, GitHub, Google Datasets, and Kaggle do not have this dataset categorization functionality, which is a plus point for the open data platforms. | + | + | - | - | + | + | - |
| Customization | The customization of functionalities provided by the OD platforms are important factor to analyze the platforms. | Available in open data portals/CMS platforms but expertise required to customize the functionality. GitHub, Kaggle and google datasets did not support customization of portals, just you can upload the datasets. | + | + | - | - | + | + | - |
| Supply and Quality of Data in each OD platforms | Supply and demand in the open government data portals are important and its individual government personnel responsibility to maintain the supply and quality of data standardized [3]. | The supply and quality are completely dependent upon the data providers. So, if open data providers are vigilant in maintaining the supply and quality of open data, they can do it on any platform. | + | + | + | + | + | + | + |

**Table 1: (*Continued*)**

| Features/ Properties | Definitions | Detailed Analysis and comments | CKAN | DKAN | GitHub | Google Datasets | ODS | Socrata | Kaggle |
|---|---|---|---|---|---|---|---|---|---|
| Demand and engagement | The demand driven approach in the open data publishing is really need of the society to solve the problems timely. | This functionality is followed by several countries which are based on CKAN, DKAN, Socrata, and openDataSoft. | + | + | - | - | + | + | - |
| Interoperability and Standardization security, privacy, and transparency | Interoperability: secure unrestricted communication of data and services among the portals through proper network channels and protocols. | Interoperability and standards are followed with different available tools and techniques such as security protocols. Semantic, syntactic, organizational, and legal interoperability are four layers of interoperable open data portal [28] [29]. | + | + | - | - | + | + | - |
| Real-time-data charts and graphs | This is very dynamic functionality, because at the same time an open data portal receives different types of data from different entities such as government, research, finance, and science. Few portals provide this functionality for specific data types and many not. Onsite analytics is interesting to implement, one can integrate the available tools such as statistics tools extension in OGD portals for quick understanding of data [19]. | Most open data uploaded to data platforms is dynamic data from different fields, with different data structures and file formats. That's why it is difficult to generate charts on dynamic and distinct-in-nature data. For each type of data, you will need to develop real-time data charts and graphs functionalities by using the customization options available in CKAN, DKAN, Socrata, and ODS. | - | - | - | - | - | - | - |
| Management of ontologies | Is the platform equipped with the mappings needed to connect the data source to the ontology and deliver the services? | Management for ontologies are provided using the governmental setup. For instance, municipalities and ministries will handle the ontologies as they wish to serve the citizens in each platform. | + | + | - | - | + | + | - |
| Trusted Data Intermediaries | Government officials are intermediaries. Most of the time trusted, but GitHub and google datasets did not confirm the trustees of intermediaries. | The trustworthiness of intermediaries is well mamaged by open data platforms such as CKAN, which is very widely used in governmental open data projects. This trust is also provided by DKAN, Socrata, and ODS, but to a lesser extent than by CKAN. GitHub has trusted data intermediaries, but not more trusted than CKAN, DKAN, Socrata, and ODS. Google Datasets did not support all of this functionality. | + | + | - | - | + | + | - |

**Table 1: (*Continued*)**

| Features/ Properties | Definitions | Detailed Analysis and comments | CKAN | DKAN | GitHub | Google Datasets | ODS | Socrata | Kaggle |
|---|---|---|---|---|---|---|---|---|---|
| Data linkage | Data linkage is the property of two or more than two portals to link or share the data among them. | Machine-to-machine and portal-to-portal communication for data linkage are available, but they are very complex and difficult to implement. CKAN, DKAN, Socrata, and ODS provide this functionality for open data based on vocabulary and metadata standards such as DCAT-AP and Berlin metadata standards. Google Datasets and GitHub don't have the machine-to-machine data linkage. | + | + | - | - | + | + | - |
| Organizational data policy strategies and risk management | The organization's data policies, such as how much data will be opened and how the data opening will be controlled, are important management factors. | Organizational policies and other factors can also be controlled by proper resources. CKAN, DKAN, OpenDataSoft, and Socrata can support all these points. | + | + | - | - | + | + | - |
| Digital Literacy | Data Literacy is important to train the skill-full stakeholders such as infomediaries, intermediaries, and end-users for better understanding the valuable data. | Different NGOS such as the World bank, European Union provide data literacy programs using the CKAN portal as a baseline. For google datasets, and GitHub less resources are provided for digital literacy. | + | + | - | - | + | + | - |
| Data privacy | PSI directive and General data protection rule(GDPR) to protect the data and open data. | This follows the PSI rules to protect the personal information of the public. Almost all the platforms support data privacy at their end. Most of the time, these systems have privacy-by-design functionalities. | + | + | + | + | + | + | + |
| Harmonization | Providing the data to end-users in a combined form from different portals and data catalogs | One can report an issue to resolve the dataset problem to the CKAN portal providers. GitHub and Google datasets cannot combine the datasets from different sources to merge them and present them to end-users. CKAN, DKAN, Socrata, and ODS are open data platforms and provide a facility for data harmonization for better data findability. | + | + | - | - | + | + | - |
| Actors' role | Different roles are defined with respect to their privileges. For instance, different roles for data providers and end-users. | Open data platforms, except Google datasets, have the same actor roles for different stakeholders. | + | + | + | - | + | + | + |
| Security protocols for sharing client data | Secure security protocols used in both API based and web-based solutions to provide secure communication over the internet. | Security protocols have been implemented on all platforms to ensure the integrity of catalogs, portals, and data. | + | + | + | + | + | + | + |

**Table 1: (*Continued*)**

| Features/ Properties | Definitions | Detailed Analysis and comments | CKAN | DKAN | GitHub | Google Datasets | ODS | Socrata | Kaggle |
|---|---|---|---|---|---|---|---|---|---|
| Latest Technologies | UX/UI, Artificial intelligence, cloud computing, blockchain, servers etc. | CKAN, DKAN, Socrata, and ODS platforms can support artificial intelligence, blockchain, servers, and the latest UX/UI functionalities for open data. The GitHub and Google datasets have static technologies. | + | + | - | - | + | + | - |
| CMS Platforms | This is very important factor, Content Management system is another software which should be available to manage the content inside the open data platforms. | A few open data portals based on other CMS platforms, such as DKAN, are built on the CKAN. CKAN did not have a built-in CMS integrated, but DKAN had an already integrated CMS. GitHub, Socrata, and OpenDataSoft provide the integrated CMS. The CMS was not integrated into Google datasets. | - | + | - | - | - | - | - |
| Web servers and backend languages | Webservers and backend languages are also important to analyze the technological commons of OD platforms. | CKAN uses Nginx and Python. DKAN uses the JavaScript framework (ReactJS). GitHub uses the JavaScript framework (ReactJS). Google Datasets uses the JavaScript framework (ReactJS). OpenDatasoft uses PHP as a backend language and to develop the server side. Socrata uses JavaScript-based backend servers. | + | + | + | + | + | + | + |
| Data formats | Multiple data formats are supported by each open data portal. The most common data formats are .pdf, .csv, .XLS, .RDF, and LOD [1]. | CKAN, DKAN, Socrata, and OpenDatasoft support the five-star data model along with several other data categories. GitHub and Google datasets don't follow the 5-star model, although they follow but not more than 3 stars. | + | + | + | + | + | + | + |
| User interfaces technologies | Most common used user interfaces technologies are bootstrap, Gatsby, Django, angular, react, and some mobile versions UX/UI. | CKAN Uses (React, Django). DKAN uses ReactJS and Gatsby. GitHub Uses the ReactJS and jQuery in common. Google Datasets are also using the ReactJS technology to develop the user interfaces. ODS uses the ReactJS, and Bootstrap based developed GUI. And Socrata uses the ReactJS, jQuery, and Bootstrap. | + | + | + | + | + | + | + |
| Policies and process | How are business, data, and IT policies created and coordinated across the data life cycle? Creation, circulation, collecting, storing, usage, and destruction? | In CKAN, DKAN, Socrata, and ODS, open data publishers can mold the platforms with respect to their policies and can develop processes accordingly. GitHub and Google Datasets are just data repositories and version control hubs, so it's difficult to manage the policies and processes. | + | + | - | - | + | + | - |

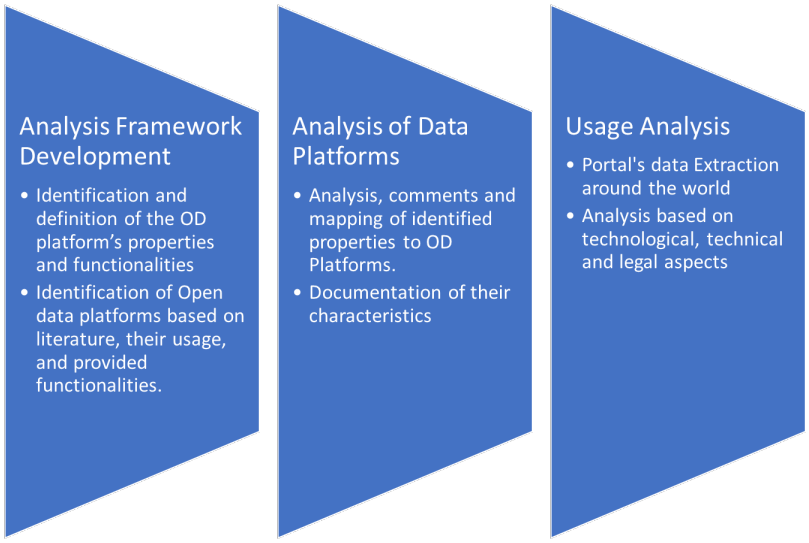


**Figure 1: A Proposed Methodology for the Analysis of Open Data Platforms**

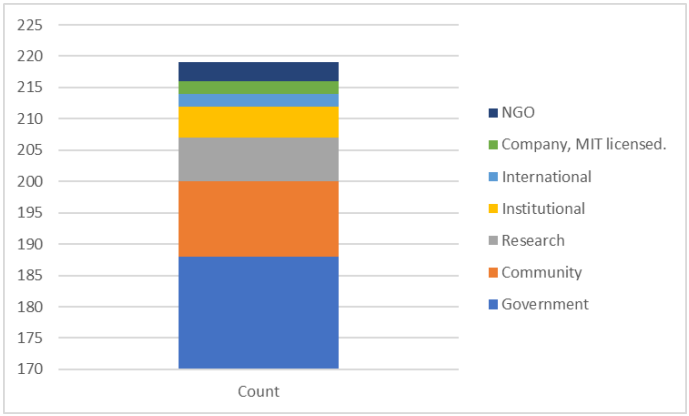


**Figure 2: Open data publisher's classification**

was introduced. This is an open license portal and is also available on GitHub repository. After getting data from Dataportals.org, descriptive statistics were used to get the information needed to evaluate the open data portal projects in different countries. We evaluated 593 open data portals from around the world based on features such as open data publishers' categories, number of portals by each country, licenses used in open data portals, importance in the world (e.g., what portals are widely used and what portals are not widely used?), type of available generators (baseline portals such as CKAN, DKAN, etc.), and most frequent words used by different data portals.

Figure 2 depicts the different open data portals' users and their count; the government has more open data portals than the other data providers. Figure 3 depicts the most commonly used open data generators which are used to develop open data portals.

## 5 DISCUSSION

A typical open data or open government data release includes the collection of data and publishing it to different open government data portals or simply on open data portals. There are several choices that exist on the internet, such as the US (data.gov), the UK (data.gov.uk), Taiwan (data.gov.tw), France (data.gouv.fr), and Singapore (data.gov.sg). These portals are famous for open government data [7]. These portals are like one-stop shops that make it easy for clients to get access to government data. This saves them the trouble of having to get information from several government agencies, offices, or websites. Open knowledge foundation provides the global open data index for different countries of the world to describe the state of the open data publication world. To date, among 94 countries Taiwan open government data publication is mature than the other countries based on OKFN global open data index.

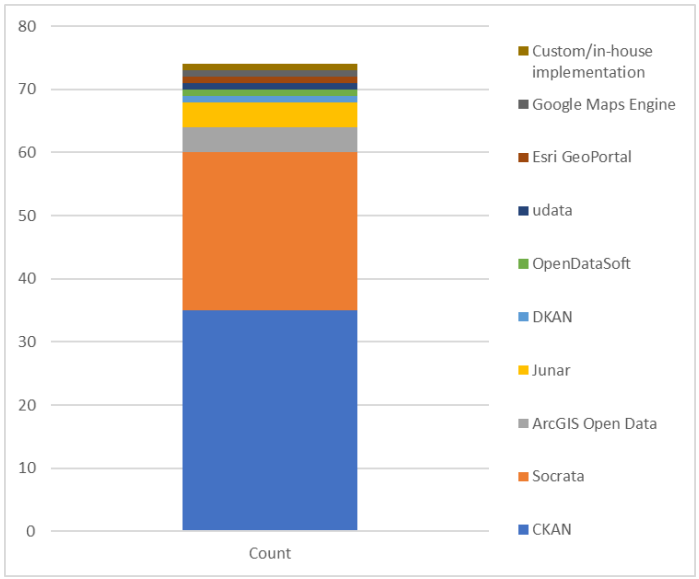


**Figure 3: Open data generators based on which other open data portals can be developed**

This index looks at the legal and technical aspects of each dataset that is available from multiple sources. The analysis framework devised in this research some solutions are suggested for the selection of open data portals. DKAN provides the integrated CMS functionality, but in CKAN it is not provided. To take advantage of CMS in DKAN, the public sector independently chooses the open data platform for their open data publishing. GitHub and Google datasets do not fulfill all the characteristics of open data platforms. Therefore, public sectors should avoid these to make data available to the public. The data must be opened with a purpose. The GitHub and Google datasets are less suitable for purposeful open data publishing because the linked data, metadata, vocabularies, and CMS properties are not available.

## 6 CONCLUSION

Open data platforms (CKAN, DKAN, ODS, and Socrata) and data repositories (Google Datasets, GitHub, and Kaggle) for publishing open data are examined in depth in this paper. The methodology developed yielded data regarding open data portals, platforms, and repositories. Each of the open data portals, platforms, and repositories is analyzed and commented upon in depth and with specificity. We have adopted the plus (+) and minus (-) signs to reflect the availability of functionality on each portal, platform, and repository, respectively. Additionally, the distinction between open data portals and metadata was defined. We developed the appropriate functions and sub-functions to investigate existing open data portals. We employed descriptive statistics to examine the global open data portal dataset and gave graphs analyzing the "top open data publishers," "widely used licenses," "top 20 nations with the most open data initiatives," and "top open data platforms that are used to create OD portals." This study demonstrated the significance of metadata, APIs, languages, data visualizations, searchability, and standards for the success of open data portals in a sustainable open data ecosystem. Data producers and data curators can choose portals, platforms, and data repositories based on their needs and intended audiences. Finally, it is determined by our methodology that CKAN and DKAN are the two best open data platforms for creating sustainable open data portals for public or private institutions, among other presented portals, platforms, and data repositories.

## REFERENCES

[1] O. H. Abdelrahman, "Open Government Data: Development, Practice, and Challenges," Open Data. London, United Kingdom IntechOpen, 2021, doi: 10.5772/intechopen.100465.
[2] B. Van Loenen *et al.*, "Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda," JeDEM - eJournal eDemocracy Open Gov., vol. 13, no. 2, pp. 1–27, 2021, doi: 10.29379/jedem.v13i2.644.
[3] Y. Charalabidis, C. Alexopoulos, and E. Loukis, "A taxonomy of open government data research areas and topics," Journal of Organizational Computing and Electronic Commerce, vol. 26, no. 1–2. pp. 41–63, 2016. doi: 10.1080/10919392.2015.1124720.
[4] Y. Tzitzikas, M. Pitikakis, G. Giakoumis, K. Varouha, and E. Karkanaki, "How Can a University Take Its First Steps in Open Data?," Commun. Comput. Inf. Sci., vol. 1355 CCIS, no. Mtsr, pp. 155–167, 2021, doi: 10.1007/978-3-030-71903-6_16.
[5] J. Nogueras-Iso, J. Lacasta, M. A. Urena-Camara, and F. J. Ariza-Lopez, "Quality of Metadata in Open Data Portals," IEEE Access, vol. 9, pp. 60364–60382, 2021, doi: 10.1109/ACCESS.2021.3073455.
[6] M. Solar, G. Concha, and L. Meijueiro, "A Model to Assess Open Government Data," Electron. Gov. Lect. Notes Comput. Sci., pp. 210–221, 2012.
[7] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," Gov. Inf. Q., vol. 32, no. 4, pp. 399–418, 2015, doi: 10.1016/j.giq.2015.07.006.
[8] European Commission, Recommendations for open data portal: From setup to sustainability. 2020. doi: 10.2830/876679.
[9] S. Kubler, J. Robert, Y. Le Traon, J. Umbrich, and S. Neumaier, "Open data portal quality comparison using AHP," ACM Int. Conf. Proceeding Ser., vol. 08-10-June, pp. 397–407, 2016, doi: 10.1145/2912160.2912167.
[10] P. Dymora, M. Mazurek, and B. Kowal, "Analysis of selected characteristics of open data inception portals in the context of smart cities iot data accessibility," WEBIST 2020 - Proc. 16th Int. Conf. Web Inf. Syst. Technol., no. Webist, pp. 67–74, 2020, doi: 10.5220/0010117600670074.
[11] R. P. Lourenço and L. Serra, "An Online Transparency for Accountability Maturity Model," pp. 35–46, 2014.
[12] L. Danneels, S. Viaene, and J. Van den Bergh, "Open data platforms: Discussing alternative knowledge epistemologies," Gov. Inf. Q., vol. 34, no. 3, pp. 365–378, 2017, doi: 10.1016/j.giq.2017.08.007.

[13] European Commission, "Towards Open Government Metadata," no. September, pp. 1–6, 2011, [Online]. Available: https://joinup.ec.europa.eu/sites/default/files/24/4c/14/towards_open_government_metadata_0.pdf
[14] D. Corsar and P. Edwards, "Challenges of Open Data Quality," J. Data Inf. Qual., vol. 9, no. 1, pp. 1–4, 2017, doi: 10.1145/3110291.
[15] J. Nogueras-Iso, H. Ochoa-Ortiz, M. Ángel Jañez, J. R. R. Viqueira, L. Po, and R. Trillo-Lado, "Automatic Publication of Open Data from OGC Services: the Use Case of TRAFAIR Project," GEOProcessing 2020, Twelfth Int. Conf. Adv. Geogr. Inf. Syst. Appl. Serv., no. c, pp. 75–80, 2020, [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=geoprocessing_2020_1_130_30086
[16] R. P. Lourenço *et al.*, "Evaluating the quality of open data portals on the national level," Gov. Inf. Q., vol. 12, no. 1, pp. 21–41, 2014, doi: 10.4067/S0718-18762017000100003.
[17] M. Lnenicka *et al.*, "Transparency of open data ecosystems in smart cities: Definition and assessment of the maturity of transparency in 22 smart cities," vol. 82, no. February, 2022.
[18] O. Bello, V. Akinwande, O. Jolayemi, and A. Ibrahim, "Open data portals in Africa: An analysis of open government data initiatives," African J. Libr. Arch. Inf. Sci., vol. 26, no. 2, pp. 97–106, 2016.
[19] M. Lnenicka and A. Nikiforova, "Transparency-by-design: What is the role of open data portals?," Telemat. Informatics, vol. 61, no. March, p. 101605, 2021, doi: 10.1016/j.tele.2021.101605.
[20] and P.-H. Y. Hsin-Chang Yang, Cathy S. Lin, "Toward Automatic Assessment of the Categorization Structure of Open Data Portals," in International Conference on Multidisciplinary Social Networks Research, 2015, pp. 372–380.
[21] DKAN, "Comparing the CKAN and DKAN," 2022. https://dkan.readthedocs.io/en/latest/introduction/dkan-ckan.html#comparing-dkan-and-ckan
[22] S. Neumaier, J. Umbrich, and A. Polleres, "Automated quality assessment of metadata across open data portals," J. Data Inf. Qual., vol. 8, no. 1, 2016, doi: 10.1145/2964909.
[23] OpenDatasoft, "OpenDataSoft Interoperability Framework," 2022. https://help.opendatasoft.com/platform/en/publishing_data/06_configuring_metadata/interoperability_metadata.html
[24] Data.Gov, "Open Data portals and Google Datasets," 2022. https://datasetsearch.research.google.com/search?query=education outcomes site%3Adata.gov&docid=L2cvMTFxOHgzdGJxZA%3D%3D
[25] A. Zuiderwijk, M. Janssen, and C. Davis, "Innovation with open data: Essential elements of open data ecosystems," Inf. Polity, vol. 19, no. 1–2, pp. 17–33, 2014, doi: 10.3233/IP-140329.
[26] C. Alexopoulos, E. Loukis, S. Mouzakitis, M. Petychakis, and Y. Charalabidis, Analysing the Characteristics of Open Government Data Sources in Greece, vol. 9, no. 3. 2018. doi: 10.1007/s13132-015-0298-8.
[27] P. F. Uhlir and P. Schröder, "Open Data for Global Science," Data Sci. J., vol. 6, no. June, pp. OD36–OD53, 2007, doi: 10.2481/dsj.6.od36.
[28] Joinup Europe, "Interoperability layers Translate," Interoperability solutions, 2022. https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers
[29] Y. Charalabidis, C. Alexopoulos, and E. Loukis, "A taxonomy of open government data research areas and topics," J. Organ. Comput. Electron. Commer., vol. 26, no. 1–2, pp. 41–63, 2016, doi: 10.1080/10919392.2015.1124720.