Application-based Framework for Analysis, Monitoring and Evaluation of National Open Data Portals

Vigan Raca[1], Goran Velinov[2], Betim Cico[3], Margita Kon-Popovska[4]
Ss Cyril and Methodius University in Skopje
Faculty of Computer Science and Engineering, Skopje, North Macedonia[1, 2, 4]
Metropolitan University of Tirana, Faculty of Computer Science and IT, Tirana, Albania[3]

***Abstract*—Open Government Data (OGD) portals are considered of significant national importance towards transparency and accountability improvement. The continuous publication of data in OGD portals introduces the need for high-quality data and the qualitative portal itself. This paper aims to address the data quality issues through a framework composed of several components aimed at measuring and monitoring the OGD portals in an automated way. Through this proposed framework, is intended to monitor and evaluate OGD quality, respectively OGD portals, and to show their progress/regress based on accumulated scores for different periods. The advantage of the proposed framework is the compatibility with any OGD Portal due to its flexibility of integration. The integration interface consists of only a few basic metrics but is necessary that almost the OGD portal possesses and can produce very compressive results. The other advantage is the possibility of extraction of collected data for further analysis and the introduction of artificial intelligence (AI) for prediction purposes to point out how the OGD portals will stand in the next period.**

Keywords—Open data; government; datasets; evaluation; portals; framework

I. Introduction

Nowadays, the trend of open data is developing at a rapid pace, while constantly increasing amounts of open data boost of development. In this regard, the role of the European Directive for using and re-using public sector data boosts the new trend towards opening up government data [1, 2]. This trend of development has gained the attention of governments and other public sector bodies for opening their data. Thus, regardless of the administrative levels, the public sector bodies are one the main publishers and holders of information for i.e. registered companies, maps [3]. The public sector data entailed the possibility for use and reuse for commercial purposes [4] while the main goal remains the increase of quality of transparency and accountability of governments [5].

Initially, in 2009 the White House promoted the Open Government Data (OGD) initiative [6] which called on all democratic states to become part of this initiative by opening their data. A few years later, in 2011, the initiative named after "Open Government Partnership", in cooperation with civil society, was established and it aims to advance and promote open data. So far, 78 countries are members of this partnership that serve more than 2 billion people to promote and strengthen the transparency and accountability of governments and increase public participation in policymaking. These institutional and global developments show that the promotion of open data has continued over the years resulting in an overall increase in the number of datasets in the disposition of citizens, scholars, businesses, and similar. Regarding terminology, in the literature exist different acronyms that differ from each other. Sometimes is referred to Open Government Data (OGD), but somewhere is used the short acronym "Open Data". When the term "Open Data" is used, it includes whatever data such: government, businesses, health, insurances, mappings, etc. But when the term includes the compound acronym as “government” or “national” it is sure that it referred to public data produced by public sector bodies [7].

The open data as a term has been addressed in early years, while the quality of open data was addressed first in 2006 by Berners–Lee is the first who published a scheme dedicated to open data quality which was based on 5 levels represented as stars [8]. This scheme is based on the quality of file format publication and rates file formats based on stars. While, data quality as a general concept is addressed in the early 90s when Wang et al discussed the dimensions for measuring data quality [9, 10].leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided.

The paper is organized as follows: Section 2 explains the methodology employed in this research; Section 3 addresses theoretical and other practices of OGD, national portals, existing frameworks, and existing portals that measure the quality of open data at the national level; Section 4 discusses the proposing and building of framework build, the scoreboard for measuring the quality of portals, the web-service for collecting information from national open data portals, data collection and classification, processing, and provision of real-time results through the dashboard, and the possibility of using of an API for data analysis by anyone or any third-party application.

II. Literature Review

The wide range of OGD may include data from various public sectors, agencies, the local level of government, ministries, universities, and many other public sectors, but all of these intersect in a portal entitled national portal or OGD

Portal [11]. The OGD portal is a national single point where public sector bodies (organizations) of the country make their data available with the purpose of strengthening transparency and integrity.

In addition, the OGD portals are a simple website interface through which they facilitate the use of published data so that citizens and other non-governmental actors can use them. The data published on these portals are usually recorded in the form of metadata organized in rows and columns containing different information depending on the government sector bodies [12].

These OGD portals have constantly changed, contributing to the needs and demands [13]. Initially, they only intended to serve as interfaces where data in the form of datasets are published, then over time and need, they have advanced, enriching themselves with other features [14]. The addition of other features based on needs has pushed forward a more efficient use of data [15]. The addition of various search filters, the provision of more information on the data producer, grouping in the form of data types such as (economy, public safety, finance, justice), etc., are some of the advancements in time. In combination with the above-mentioned functionalities, those OGD portals have developed their application programming interfaces (APIs) to allow the consumption and query of the data by the third-part application in an automated manner [15, 16]. The API is a software intermediary that allows two applications to talk to each other. The availability of this feature has greatly facilitated the work, where access to the resources of OGD portals can be automatically provided for the use of published data through a third application completely automatic.

The availability of APIs, especially for government open data portals, has given them many opportunities in addition to the automated use of resources, and also opened the way for analysis and quality measurement of portals, and publishing data [17]. In this context, several portals have been developed that aim to monitor the quality of open data portals at the national level including the global open data index [18], open data watch [19], open data barometer [20] etc. Compared with mentioned portals above, there is used a different approach proposing a new evaluation model. This model in principle is based on those portals but, unlike them, the proposed framework monitors and evaluates them in real-time by providing the following information: number of datasets, organizations, groups, tags, licenses, and type of datasheet formats.

Another characteristic of the proposed framework is that it uses a benchmark based on a multidimensional model that makes it a perfect combination. A framework in the context of data quality is a kind of assessment tool that helps to measure the data quality of organizations aimed to improve the quality. This combination uses file formats of published datasets and information about these datasets. All this nomenclature is defined as a framework model which easily interacts and expands with various national open data portals by connecting to their APIs. Initially will be applied to open data portals in six western Balkan countries [1] (Albania, Bosna and Herzegovina, Kosovo, Montenegro, North Macedonia, and Serbia).

III. Related Work

This research paper analyzes and discusses the existing approaches that have been proposed and adopted for monitoring and evaluation of OGD quality. In addition, it discusses the actual frameworks proposed as well as current tools and portals for evaluation of OGD data quality aimed to propose the development of a new evaluation model framework employing qualitative and quantitative approaches combined and interacted to provide compressive evaluation results. This perfect combination is conceptualized as a framework model consisting of several other components discussed in further sections.

A. Existing Approaches and Tools

Open government data portals have continuously developed and advanced, particularly in developed countries where this revolution has initially begun [21]. Numerous needs for access to data have also influenced the further development and advancement of national OGD portals. This development and advancement include the improvement of data quality and quality of portals as well.

When it comes to quality, so far various aspects of data quality from the definition, types, dimensions, techniques, strategies, and multidimensional proposals have circulated [22, 23]. In this respect, different frameworks for measuring data quality have been developed. Since the purpose of research is based first on designing and conceptualizing a framework for measuring the quality of open government data, for this reason, different frameworks have been analyzed. According to Maurino et al. quality is related to dataset level, but the evaluation is performed at the portal level by aggregating the values computed on each dataset [24].

In addition to current frameworks developed, continuous progress has been made by building portals with the aim of monitoring and measuring OGD quality. In this context, different portals are available today such: open data index, open data barometer, open watch data, open data EU, etc. Some of them measure only OGD quality but some others monitor also the number of resources published by OGD portals in the context of datasets, organizations, licenses, etc.

Each of these portals uses its methodology based on the framework through which the quality of open data is measured or monitored. Based on the analysis performed, the frameworks that use classification of datasets based on the profile, for example (economics, judiciary, finance, law enforcement, statistics, health, etc.), as well as for each field questionnaires have been applied on publication of data such as: are they licensed? Are they in machine (readable) format, are they available? Can this data be manipulated by asking for sprawl? How often are the data published? etc. So, these types of calculations are used in the open data index frameworks, open data barometer using scores for each part of the

[1] Countries are listed based on alphabetic order

evaluation, and deriving the average result for each national OGD portal by ranking portals based on countries.

The following sections will discuss the features of each by comparing them.

The Open Data Barometer (OBD) is an advanced system that evaluates government open data. This system relies on score listing countries based on questionnaires on policies, implementation, and impact of data initiatives as well as openness assessment built from 14 data types for each country. So, this system evaluates datasets through a review process in 14 different areas such as legislation, transport, health, crime, procurement, etc.

Open Data Index (ODI) is a crowd-sourced indicator dedicated to the openness of datasets, which was founded by the Open Knowledge Foundation. Furthermore, the information on the datasets is collected based on the open data census, creating an index for each country, and scoring them by undergoing a process by going through 9 attributes based on the Open Definition. This rating system has an ideal value of 100 (maximum) for each attribute. The maximum weight is 30 points that are dedicated if the database is license open. While datasets that are not accessible have a score of 0.

Open Data Monitor, unlike the two systems mentioned above, is another similar system, so it has almost the same purpose, where in addition to evaluating the quality of OGD, it also monitors the available resources, presented in the visual form using the most innovative technologies. This system has a framework that is based on the dataset and metadata from OGD resources. The process of gathering the necessary information from national open data portals keeps it in a structured form for further processing. In this context, it uses analytical and visual methods to be user-friendlier for users and to give results in visual forms, unlike others that display in statistical form. Another value of this system is that it enables comparison between countries to also see visually the results of each. The platform uses the exposed APIs of the open data portals of the national level of the EU members. In terms of functions, this system offers a range of analytical functions such as comparison of public bodies (national/local), metadata quality, selection of catalogs for different fields, license information, published dataset formats, last updated, percentages, etc. All of these are characterized based on qualitative and quantitative methods. Quantitative refers to the quantity (number of resources) available, while qualitative includes the analytical functions mentioned above.

So, if compared to open data index and open data barometer, the open data monitor system is not global, but it is dedicated only to EU countries, it also uses analytical tools and displays data in a very attractive way using tools for data visualization. It is important to note that the latter (open monitor data) in contrast shows information only until 2015, while the open data index and open data barometer display data based on the current global situation. Since our research aims to measure the quality of portals and monitor them, the analysis of existing frameworks will help build a multi-functional framework that performs quality measurements and monitors at very frequent periods each week.

In addition to analyzing existing OGD evaluation portals, few scientific articles have been reviewed related to benchmark frameworks. According to Renta Machova et al, they have used other data sources and different information they have collected [25]. Also, the framework proposed by them consists of more than 20 metrics that complicate the process of evaluation due to the high probability of changing and updating portal APIs. Also, is not sure if all portals possess those metrics information. While according to Antonio Vetro et al they also have proposed a framework that is based on two dimensions consisting of several metrics of around dataset and other metrics for evaluating data quality of data within the dataset [26]. So, besides the information about the publication of the dataset this framework measures the quality of data inside the dataset (records). It is very valuable, but there is not implemented in any machine for performing an automatic evaluation, but they have applied it manually by checking each portal separately. Referring to Peter Parycek et al, they have developed a method for evaluation of OGD that is applied in the city of Vienna [27].

This method is based on surveys prepared and sent to respondents. A framework proposed by them is more suitable for regulating data publication and provides recommendations on how to publish high qualitative data than evaluation of existing data published.

Therefore, compared with those frameworks, the proposed framework uses a different approach, is can be easily scalable and can be implemented and integrated into application in a very easy way. Initially, the information target to collect is very basic, so most of the portals possess this information, and the probability to change the APIs or missing this information is very low. This empowers the proposed framework because with only a few metrics will be possible to evaluate and monitor the OGD portals at any time. Based on the discussion about existing frameworks, something different will be proposed that will fit any portal, be well integrated, and works independently with no need for human intervention.

Once collecting and storing of data into the database will be performed from target portals, another feature of the proposed framework is the ability to share data through the API to anyone that may be interested in further development, analysis, or using any third-party application, it is possible only by integration or API provided.

IV. Research Methodology and Methods

This research utilizes a qualitative approach applying mixed methods that combine analytical rigor and data gathering, as well as monitoring and evaluation.

Intention to build a framework model for monitoring and evaluation of the quality of OGD portals based on a scoreboard that displays the scores for each dimension metric is based on specific methodology In this respect, the proposal for the build-up of the framework model is divided into several phases as follows:

Phase I. Analysis of OGD National portals, their review, and general evaluation of whether these portals provide APIs as a prerequisite for further monitoring and evaluation.

Phase II. Identify resources within selected portals, what APIs they offer, and find common denominators to ensure that all selected portals meet each parameter set.

Phase III. Proposing of benchmark framework design based on analysis of existing OGD portals. This proposed framework will perform monitoring of OGD portals and evaluate quality.

The proposed framework is designed taking into account the following steps:

1) Analysis of existing frameworks for measuring the quality of open government national portals.

2) Targeting data sources for application of the framework.

3) Defining the dimensions to be applied.

4) Defining metrics for each dimension.

Phase IV. Once the framework is defined, it remains to be integrated into the framework model, which consists of several components, starting initially with the first component of the web service that will have several roles:

1) Integration/interconnection with APIs of targeted portals in this research.

2) Collection of data required for the Framework definition in Phase III.

Phase V. Once the necessary data has been provided, there is now another phase, which deals with the processing of this data, the analysis, and displaying of the data. Furthermore, within this phase, there will be some processes as follows:

1) Data processing through validation and cleaning process.

2) Application of the application-level framework for measuring the quality of the processed results.

3) Display results in the Web interface (dashboard) in real-time for OGD portals that have been selected.

Phase VI. Developing an API and making it available to anyone. It will provide the data collected by saving time and work because there is not necessary to connect each portal APIs for getting data, since this data already exists but will be shared through an API.

V. Analysis of Open Government Data Portals

The main purpose of this research is to propose and build a system or tool that will consist of many components defined as a framework model that monitors and measures the quality of national open data portals in an automated way. Therefore, a basic prerequisite for building a framework is the definition of basic needs. Thus, first, it is necessary to analyze portals that will be the target of monitoring and evaluation, which include Western Balkans national open government data portals, (Albania, Bosnia and Herzegovina, Kosovo, North Macedonia, Montenegro, and Serbia) [31-36].

Various analyses over the OGD national portals can be applied using different criteria [28]. In this respect, five criteria analysis will be used for designing a benchmark framework and these criteria include the following questions: 1. Does the portal provide and have an available API for connection? 2. Does the portal provide the datasets for each publisher? 3. Does the portal provide the file format types published for each dataset? 4. Does the portal provide the published dates and last updates of datasets published? 5. Does the portal provide the license used for each dataset?

These five criteria are the fundamental precondition for the selection of government portals for further monitoring and evaluation.

Since the term “a framework” has been used everywhere in this research, it means that will be applied to only a few OGD portals, but with the potential to be applied to other OGD portals. For the building of the framework, initially, some preconditions have been defined starting with information that should be collected because for sure that designing of the framework will be based on such information. In this regard, the analysis of available information will be performed, respectively what information the national OGD portals offer and if all OGD portals share this information.

The following table shows the necessary information and information identified in each government portal analyzed. The same information will also be used for designing the framework model.

The data defined in Table I, in addition, to building the framework model will assist the web service how to know what data to collect from the portals.

Moreover, the analysis depicts the lack of proper organization, so no standard has been used compared to the open government portals of other EU member states. Even though these portals support more than one language, the mother tongue of the countries dominates. For example, when a dataset or resource is published, the same should be published in at least another international language (English); however, these publications are mainly done in the mother tongue language.

The analysis also highlights the inadequate standards of file formats used for data publishing. In this context, it emphasizes that portals also use formats of published metadata that are out of range according to open data standards. In addition, there are identified about 20 types of dataset file-formats including formats that have used compression (.zip, .rar) that are out of any criteria. For instance, Cyrillic letters are out of any standard for extensions or international standards, yet they are used.

TABLE I. The Target of Information to be Collected

	Target Data	**Types of Information**
Quantitative	Datasets	Number of Datasets
	Publishers	Number of Organizations
	Groups	Number of Groups
	Licenses	Number of Licenses
Qualitative	Datasets	Dataset File Format Types
	Publishers	Publisher's Names
	Groups	Public Sectors Bodies
	Licenses	Types of Licenses

There are other cases where the publication date is outside the standard or they show no information concerning the data published. This context complicates the qualitative evaluation of the data. Thus, it was necessary to use techniques for equivalence of this data, to be able to evaluate the data. Lack of up-to-date dataset descriptions (What database is it? Who owns it?). In addition, these are only a few of the findings that have been identified during the analysis of portals and which at the same time have complicated and challenged the measurement of data quality. Then the lack of the type of licenses, under what license the published data operate, the lack of frequent updating, or the date of the publication itself, so all these are some of the findings during the analysis phase of the portals.

Therefore, this leads to the need to build a mechanism that would fix these problems during the publication phase where it would ensure the high quality of the published metadata but also the portal itself that serves that data.

Therefore, this is the reason why this paper, in addition to measuring the quality of data, also measures the quality of the national open data portals themselves.

In addition to these findings, the possibilities offered by these national open data portals for the automatic consumption of data that supports third-party applications. In this aspect, is almost clear that each portal provides the possibility of consuming data through APIs, so there is an API available, while each has its limits in the context of what they offer. CKAN based API mainly dominates, but some are based on DKAN. This also fulfills the primary condition, the collection of initial data. Although the documentation on how to consume these APIs, exists in their mother portals, even in the national portals they have published additional documentation, this also facilitates the use of the method for data collection.

CKAN [2] is the world's leading open-source data portal platform. It makes easy publishing, sharing, and working with data. In addition, it is a kind of data management system that provides a powerful platform for cataloging, storing, and accessing datasets with a rich front-end, full API (for both data and catalog), visualization tools, and more [29].

DKAN [3] is a Drupal-based open data portal based on CKAN, the first widely adopted open-source open data portal software. CKAN stands for Comprehensive Knowledge Archive Network [30].

Table II presents the APIs of the Western Balkan countries, where this framework model will be applied.

TABLE II. NATIONAL OGD PORTALS APIS AND URLS

Country	OGD National Portal URL	API Model
Albania	https://opendata.gov.al/	CKAN
Bosna and Herzegovina	https://opendata.ba	DKAN
Kosovo	https://opendata.rks-gov.net/	CKAN
Montenegro	https://data.gov.me	CKAN*
North Macedonia	https://data.gov.mk/	CKAN
Serbia	htttps://data.gov.rs	CKAN*

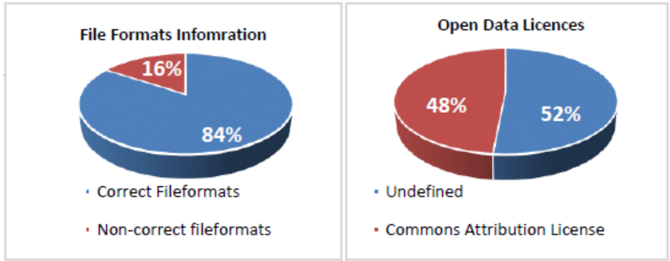


Fig. 1. (a) Analysis of File Formats Published; (b) Analysis of Licenses.

CAKAN* means that national portals have developed their API for providing information but is based on the CAKAN model. It is important to note that it is well explained with detailed information on how to use API. Apart from the investigation of APIs from the Western Balkans OGD national portals, there is analyzed the target information defined in Table I, with the attention of possible interventions on quality improvement if needed. Therefore, Fig. 1 shows the analyzed information for (a) file formats and (b) open licenses.

B. Authors and Affiliations

Correct file formats include formats published in PDF, DOC, XSL, XSLX, CSV, HTML, XML and JSON. Non-correct means the other types. While regarding licenses, the Open Data Commons Attribution License is a license agreement intended to allow users to freely share, modify, and use this Database subject only to the attribution requirements set out in Section 4 [4](Open Data Commons Attribution License ODC-By).

VI. PROPOSING OF FRAMEWORK

The proposal for the building of the framework model consists of several components and each component has its role. Because the analysis performed over OGD Portals, it precisely defines all the flaws and what information is available and can be collected from the portals for the framework model to perform its function. Fig. 2 presents all block components.

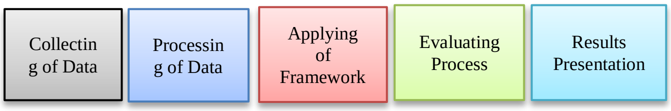


Fig. 2. The Components of Proposed Model.

First, this component means building a web service that will collect data from the OGD national portals that are targets for monitoring and evaluation. Second, this component will do data processing through insertion into the local database and data preparation. Third, conceptualization and designing of the framework. Fourth, implementation of a framework into the software application, and the fifth component is dedicated to results showing in a dashboard. The following sections explain the role and function of each component separately.

A. Collecting of Data

The first step to secure the information mentioned in the sections above is to develop a web service that will be able to

[2] CKAN, (www.ckan.org/about/).

[3] DKAN Open Data Platform (ww.getdkan.org)

[4] https://opendatacommons.org/licenses/

communicate with OGD national portals of the Western Balkan countries using the APIs available. Depending on the need, the web service can monitor and collect information on a daily, weekly, or monthly basis, made possible through configuration. Since the government data portals are open, during the analysis they did not show that they publish a large number of resources daily, the web service developed can run on a schedule on a daily, weekly basis, or monthly basis. But it depends on needs.

For the development of web services, the Microsoft .NET platform is used. The reason for using the Microsoft platform is due to practical experience and not for any other reason, Yet, this could be developed using other platforms such as java, python, visual basic, etc. Regarding the functionality of the web service, it is compiled to run as a console application. It means that the web service will not be running all the time but is configurable to run on schedule. It depends on how frequently portals publish resources or how often is needed to have refreshed results.

This process is called "Snapshot". Let's say the last snapshot is (01/08/2021 12:33), which means that the monitoring and evaluation process was performed on the data collected by (01/08/2021 12:33), indirectly the last run of web-service for collecting information was at 01/08/2021 12:33. Moreover, snapshots can be created on each day, which indirectly means that the web service will run each day at a specific date and time, respectively based on the schedule configuration. Running of web-service is not a process that only establishes connections to respective APIs, but on the other hand, it collects information. Fig. 3 shows the information that the web service will collect.

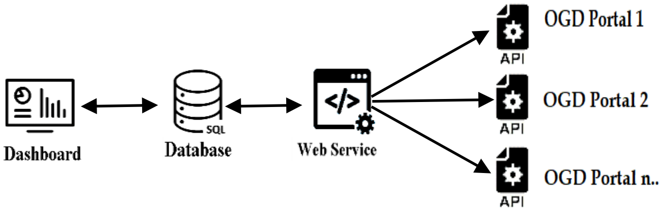


Fig. 3. Collecting of Types of Information.

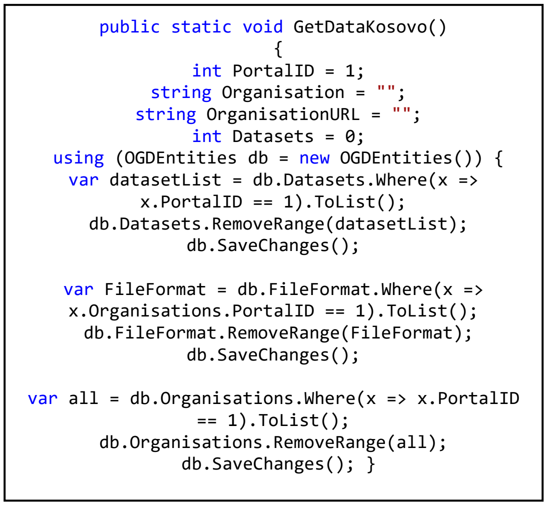
B. *Processing Data*

After collecting the targeted data, there is another extra-independent process called data processing. Within this process, three other sub-processes are performed (insertion, validation, and data cleansing). These are very important to prepare the data for further evaluation and analysis.

1) Data inserting: For the collected data to be always accessible and available, there is necessary to be stored somewhere. For this purpose, a small, but very useful database is developed. This database will also be used for other purposes such: processing of information, analysis, statistics, and showing other results. The database is developed in Microsoft SQL Server 2016 (Express Edition) but does not limit the possibilities of using other platforms. There is no reason why this platform is used, other than experience and cost-free.

As mentioned above, the analyzed portals have relatively low-quality data, so the preparation of data is inevitable, to increase the quality and enable a more accurate assessment. Given that, the web service does not do this job, it will only collect data but another sub-process will be needed (Data Preparation).

The insertion process is based on an algorithm built for this purpose, which collects the resources defined in Fig. 2 and stores them in this database based on the logic explained in Fig. 4.



```
public static void GetDataKosovo()
                {
        int PortalID = 1;
     string Organisation = "";
    string OrganisationURL = "";
        int Datasets = 0;
using (OGDEntities db = new OGDEntities()) {
 var datasetList = db.Datasets.Where(x =>
      x.PortalID == 1).ToList();
  db.Datasets.RemoveRange(datasetList);
         db.SaveChanges();

 var FileFormat = db.FileFormat.Where(x =>
 x.Organisations.PortalID == 1).ToList();
  db.FileFormat.RemoveRange(FileFormat);
         db.SaveChanges();

var all = db.Organisations.Where(x => x.PortalID
          == 1).ToList();
   db.Organisations.RemoveRange(all);
        db.SaveChanges(); }
```

Fig. 4. Collecting and Inserting of Data.

2) Data correction: Once all the data defined above have been successfully collected and stored into the database, these data will be subject to the validation process for making it ready for further processes. After the data review process, it identifies that a few data should be corrected and validated. In this matter, have been set some criteria's for correcting and validating data and figured out which data should be subject to validation. The fields of data that should be validated and corrected include publication date-time or last updates, name of licenses, and file format extensions. The next paragraph will discuss the problems of poor data quality gathered. First, there is checked for formats (extensions) of datasets, and in initial findings figure out that about 16% of them are out of range of open data standards (open data standard). Many file extensions were published in the wrong format for i.e. instead of JSON is used "GEJSON" or in the Cyrillic Alphabet in the native language is written. Second, we applied the validation to the date/time and updated dates of datasets published.

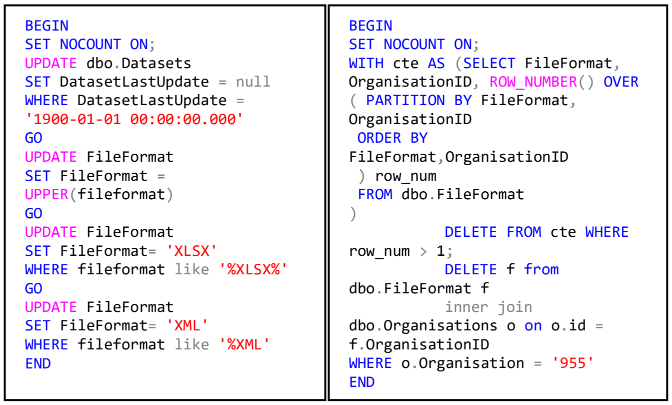
This is because each portal has used its time format such as (2020-01-10, 20-Feb-21 or Jan-03-2021 or May / 12/2021), therefore based on these facts is needed the validation of time, turning them into an acceptable standard YYY / MM / DD. The same was done with licenses, because in the findings during the analysis, about 52% of licenses were undefined, or not in the standards defined by open knowledge (Licenses - Open Data Commons: legal tools for open data).

After completing the process of data correction and validation, the precondition for data comparison between portals has been completed, but there are some issues with unnecessary data and the existence of numerous null values.

3) Data cleansing: In addition to the validation process, which means the transformation of data from one data to another without spoiling its character, we have also applied data cleansing. This sub-process has been very adequate, initially to remove useless information about the framework. Web-service collects different information, depending on how they are published, but does not use any method that validates or corrects them during the inserting process because it would complicate the whole process. Therefore, after collecting and inserting data into the database, we have applied a procedure that cleans by removing unnecessary data. During the data review process, we are faced with a lot of null values, lack of a standard for the naming of datasets and organizations for i.e. some dataset names and organization names have used the underline or line between words, some others have used spaces between, some other have used short letters for every publication made, etc.

First, removing of "null" values, and then unnecessary spaces between the names of organizations and datasets. Second, using the operations "Trim" and "Upper" for formatting the file formats extensions to have a standard and to increase evaluation accuracy.

Moreover, both sub-processes (data correction and data validation) are implemented in stored procedures of the database and both of them are triggered every time after the new snapshot. It means that every time the web service is run and after collected data is successfully inserted into the database, then those stored procedures will be triggered (executed). Fig. 5 shows two examples of data cleaning and data validation used by the framework.



```
BEGIN
SET NOCOUNT ON;
UPDATE dbo.Datasets
SET DatasetLastUpdate = null
WHERE DatasetLastUpdate =
'1900-01-01 00:00:00.000'
GO
UPDATE FileFormat
SET FileFormat =
UPPER(fileformat)
GO
UPDATE FileFormat
SET FileFormat= 'XLSX'
WHERE fileformat like '%XLSX%'
GO
UPDATE FileFormat
SET FileFormat= 'XML'
WHERE fileformat like '%XML'
END
```

```
BEGIN
SET NOCOUNT ON;
WITH cte AS (SELECT FileFormat,
OrganisationID, ROW_NUMBER() OVER
( PARTITION BY FileFormat,
OrganisationID
 ORDER BY
FileFormat,OrganisationID
 ) row_num
 FROM dbo.FileFormat
)
           DELETE FROM cte WHERE
row_num > 1;
           DELETE f from
dbo.FileFormat f
           inner join
dbo.Organisations o on o.id =
f.OrganisationID
WHERE o.Organisation = '955'
END
```

Fig. 5. Examples of Data Validation and Data Cleansing.

C. Conceptual Design of Framework

Once the data preparation process has been completed, i.e., the data served is ready for further processing, this paves the way for the design or development of the framework. Where in the state of art, we had argued quite well, the existing frameworks, showing the features and characteristics of each. Now designing the framework is considered the main work that also gives the main value of research. The proposed framework will be two-dimensional, which means it performs two different functions: monitoring national portals and measuring their quality. Therefore, for this purpose, will be used two indicators: Qualitative Indicator and Quantitative Indicator.

1) Quantitative indicator: This indicator is based on the quantitative methodology, which will have a monitoring role, which will monitor portals that count publishers, datasets, licenses, and group datasets based on the file format that is published based on the 5-star scheme. Furthermore, in Table III, we present the metrics that this indicator uses:

TABLE III. METRICS OF OPENNESS INDICATOR (QUANTITATIVE)

Scores	**Description**	**Key**
★	whatever format pdf, image, doc, text	Open License
★★	machine-readable structured format .xsl, .xlsx	Readable
★★★	non-proprietary structured format, csv	Open Format
★★★★	RDF Standards xml, html, json	URL
★★★★★	Linked to other data sources	Linked Data

Each dataset is subject to the process of evaluation, evaluation based on the file format that has been published. Observations are used to give (scores) for each metric that will be applied over datasets.

2) Qualitative indicator: Unlike the quantitative indicator, here it will do processing of information that characterizes a dataset. In this aspect, it is characterized by four main features of the dataset which we estimate affect their quality as well as the portal itself. Table IV presents the metrics used by this indicator for evaluating datasets giving it a score.

TABLE IV. METRICS OF DATASET INDICATOR (QUALITATIVE)

Observation	**Metric**	**Description**
[DAV]	Availability	Dataset is available in the portal
[DAC]	Accessibility	Dataset can be freely downloaded
[DAD]	Discoverability	Dataset is searchable (query data)
[DAT]	Timeless	Dataset is up to date

D. Assessment and Evaluation

The Framework mentioned in the above session, consisting of two indicators (quantitative and qualitative), will be applied to the framework, practically different functions translated into SQL will be used, which will produce the right results. Practically, as soon as the process of importing or inserting data from the web service in the Database has been completed, as well as the process of validation, correction, and cleaning, the data are ready for evaluation. In addition to these processes, another pre-evaluation process will be data modeling so that the application of the framework is easier.

In this regard, have been created and used several Database Views in particular for monitoring portals in quantitative terms, how many databases are available, and how many organizations publish data. Dynamic Views have been used to reflect the results dynamically on every update that may happen. This is shown in Fig. 6.

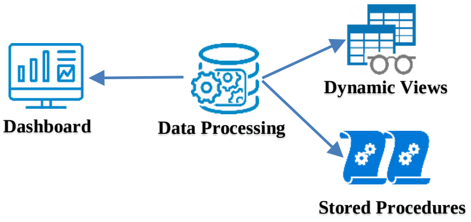
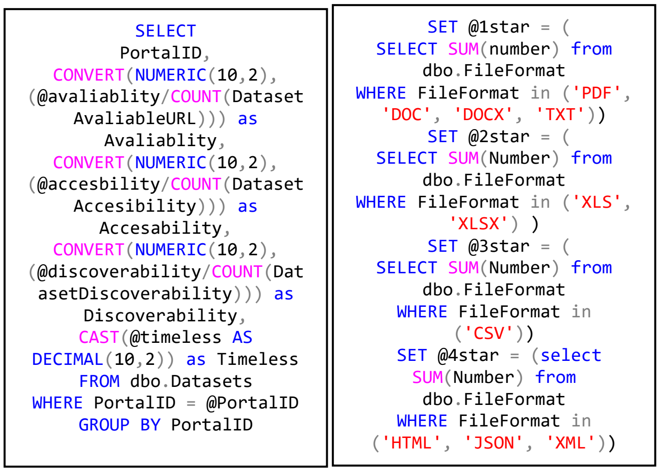


Fig. 6. Data Processing and Evaluation.

The whole monitoring process is based on VIEWs, so this is the reason for using Dynamic rather than static Views due to the changes of results dynamically based on last updates. Moreover, the proposed framework is divided into several segments; each segment uses a stored procedure, so there is no technical possibility for the whole framework to be incorporated in one stored procedure. This is due to a lot of calculations that have to be done for providing evaluation results. However, these segments depend on the metrics, which means that each metric is a stored procedure in itself.

These stored procedures will be used by the front-end part (Dashboard) illustrated in Fig. 6 quantitative indicator part, then the stored procedure of grouping and evaluating the datasets based on the file formats based on So, depending on the evaluation required, it will call and execute a specific stored procedure. Let's say, if the interest is in the 5-star scheme of Berners-Lee, will be executed and the result will be returned. However, if the interest is in the qualitative indicator, then the procedures for each metric will make their calculations and will yield the result, or both indicators, for each metric we measure by giving points (scores). For example, the data openness evaluation, which is based on the 5-star scheme, measures how open the datasets are based on the publication formats, here the evaluation is done from 1 to 5. Finally, the average for the portal. As for the qualitative indicator, this is based on the quality of the dataset based on the surrounding factors explained in Table VI.

In contrast, here the ideal or maximum value is 1 per metric, while 5 maximum values if a dataset contains all metrics. Here too a series of calculations are performed in the background, where in addition to deriving the average for each dataset that is subject to evaluation, the general average per portal is also derived. This is very important in the analytical and comparative part between open data portals. Fig. 7 presents some parts of the code for specific metrics.



```
SELECT
PortalID,
CONVERT(NUMERIC(10,2),
(@avaliablity/COUNT(Dataset
AvaliableURL))) as
Avaliablity,
CONVERT(NUMERIC(10,2),
(@accesbility/COUNT(Dataset
Accesibility))) as
Accesability,
CONVERT(NUMERIC(10,2),
(@discoverability/COUNT(Dat
asetDiscoverability))) as
Discoverability,
CAST(@timeless AS
DECIMAL(10,2)) as Timeless
FROM dbo.Datasets
WHERE PortalID = @PortalID
GROUP BY PortalID
```

```
SET @1star = (
SELECT SUM(number) from
dbo.FileFormat
WHERE FileFormat in ('PDF',
'DOC', 'DOCX', 'TXT'))
SET @2star = (
SELECT SUM(Number) from
dbo.FileFormat
WHERE FileFormat in ('XLS',
'XLSX') )
SET @3star = (
SELECT SUM(Number) from
dbo.FileFormat
WHERE FileFormat in
('CSV'))
SET @4star = (select
SUM(Number) from
dbo.FileFormat
WHERE FileFormat in
('HTML', 'JSON', 'XML'))
```

Fig. 7. Openness and Dataset Evaluations (SQL Code).

E. Presentation of Results

All calculations discussed in the section above, based on different scenarios are performed on the database level, so the results were displayed by SQL. To present these results in the right visual form, it was necessary to create a public portal in the form of interactive dashboards.

Therefore, for this purpose, a web application is developed using the .NET platform, which displays the monitoring results and measures the quality of government open data portals. The following section reflects some of the results obtained from the calculations performed to give the value final framework model, starting from the front dashboard that displays the monitoring results (see Fig. 10 in Appendix). While in Fig. 11 (Appendix), are presented the displayed results on the openness dashboard, which presents how open the portals are.

The evaluation was performed using calculations based on the openness dimension i.e., file formats of the datasets, grouping, and counting them.

The results are based on the quantitative indicator, as they do not use any other measuring feature of the dataset except the statistical one, i.e. counting and grouping. Furthermore, Table V shows the results of the qualitative indicator, i.e. the quality of the datasets, ranking the portals based following indicator metrics (Availability, Accessibility, Discoverability, and Timeless) shown in Table V.

TABLE V. DATASET INDICATOR AVERAGES (QUALITATIVE)

Country	Availa.	Access.	Discov.	Timeless
Albania	1	1	0.48	0.41
Bosna and Herzegovina	1	0.98	0	0
Kosovo	1	1	1	0.17
Montenegro	1	1	1	0.33
North Macedonia	1	1	0.81	0.25
Serbia	1	1	0	0.26

Regarding the results expressed in Fig. 8 and 9 in the background, a series of calculations are performed, but very important to show the countries' averages.

According to a mathematical point of view, for evaluating and measuring the averages of openness, the calculation is formulated using the following formula:

$$\gamma = \frac{\sum(1\ star)*1 + \sum(2\ star)*2+\sum(3\ star)*3+\sum(4\ star)*4+\sum(5\ star)*5}{\sum Total\ datasets} \qquad (1)$$

This equation calculates the average of how open the governments are by adding the whole number of datasets rated with 1 star, then with 2 stars, so on up to 5 and proportional to the total number of datasets published for the portal. This formula is applied for cases when a dataset is published in only one format.

In addition, during the analysis of OGD national portals, this research finds out that some organizations (publishers) publish their datasets in multiple formats, for i.e. "Agency of Statistics" have published two datasets, in two file formats (CSV and JSON), while the dataset remained the same because it has the same unique ID and the same name. So, for situations like that, the formula above (1) does not promise the accuracy of results, because it calculates the total number of datasets and does not check and find out if the same dataset is published in multiple file formats. Thus, for this reason, a new approach for defining datasets published in multiple file-formats has been used.

This new approach is based on two levels of evaluation, first identification and then evaluation. Table VI illustrates this situation.

TABLE VI. IDENTIFICATION OF MULTIPLE FORMAT DATASETS

Datasets			★	★★	★★★	★★★★	★★★★★	Total
Level I	Publisher	Dataset1	0	0	0	x		x
		Dataset2	0	y	0	0		y
		Dataset n			x	y		x + y
Level II	Publisher	Dataset 1						x
		Dataset 2						y
		Dataset n						y

Table VI shows that two levels of classification have been used; first, it makes classification of datasets based on file formats and counts the total number of datasets per organization (publisher). Then, after the first level is performed, the second level identifies if any of the datasets are published in multiple file format and counts only the number of higher file formats as total by removing from the calculation of other formats published.

Referring to Table VI, in the first round of calculation "Dataset n", has multiple values (x+y), while in the second round, is identified that this dataset.

$$\delta = \frac{\sum H\ (n\ star)*n}{\sum H\ Datasets} \qquad (2)$$

H – means the highest Star of the dataset.

The final equation for generating the total average of result will be:

$$f(x) = \gamma + \delta \qquad (3)$$

f(x) – is the function of calculating the overall average of openness calculation.

Based on this function, Fig. 8 shows the averages of evaluation of OGD nation portals. Results have been grouped on monthly basis to show progress.

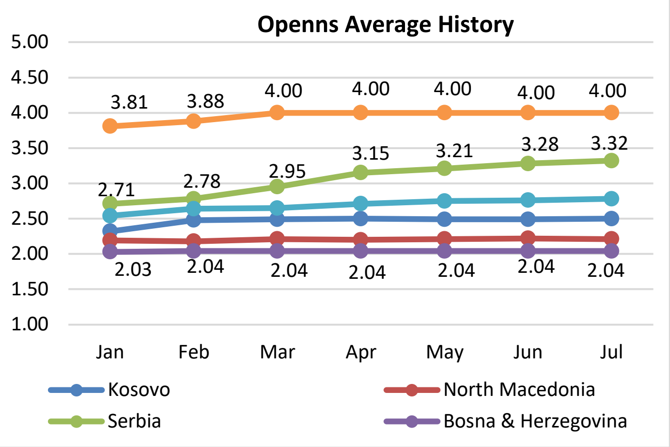


Fig. 8. Openness Averages.

In addition, the calculation of dataset quality is based on formula (4). It calculates the total average per OGD portal, respectively, it sums all the values obtained per metric in proportion to the total number of metrics used.

$$\lambda = \frac{\sum(Avaliab.) + \sum(Access.)+\sum(Discover.)+\sum(Timelss)+ \sum(n..)}{\sum(Metrics)} \qquad (4)$$

Fig. 9 shows the results produced by this formula, which is applied in the background of the application, respectively in the database implemented through SQL functions. The highest value is 1 and the lowest is 0.

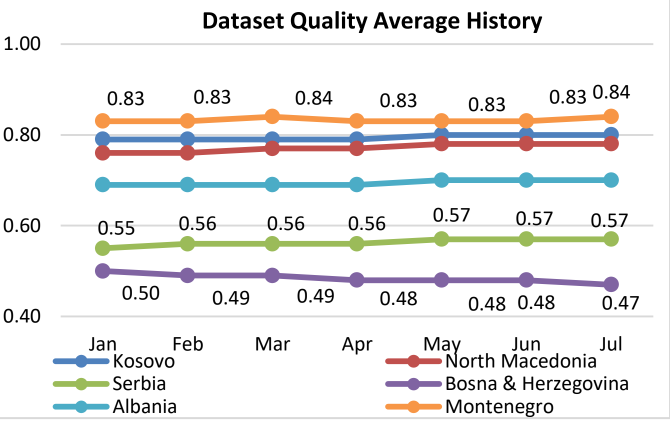


Fig. 9. Dataset Quality Averages.

In addition, the application also generates statistics, where all the results expressed in graphs, are summarized using a statistics dashboard. Statistics may change in the meantime or after each run of the web service because the data will be refreshed. All this is done automatically without the need for the human factor to intervene.

Moreover, all the data collected by the portals through the web service, after being subjected to the process of validation and clearance, can be accessible to anyone who needs this data. For this purpose, there is necessary to make available an API, which upon request returns the basic results that the web service collects such: datasets, organizations, licenses, file format types. All this is organized through a JSON API, where depending on the request, i.e. for which national portal they are required, it also returns the data. This is made available, to provide data for call part application, or anyone who needs for educational, scientific, or business purposes to have the data ready without having to develop any additional web services that take from the portals of the western Balkan countries.

VII. Result and Discussion

Through this proposed framework model is intended to monitor and evaluate OGD quality, respectively OGD portals constantly or at any time with no need for human input. An additional value of the proposed framework is that it will have the ability to show the progress/regress made by each OGD national which has been subject to monitoring and evaluation and scored in different periods.

Because of data possessed through the data collection component (web-service), the proposed framework model can show results at any time but it also can be configured to run on schedule on weekly basis or monthly basis depending on needs. Storing of evaluation history scores for each OGD Portal and visualization of results through the graphs about the progress or regress of countries gives another value to this framework. In addition to monitoring and evaluation affinities, the proposed model shares the API that will be available to the wider community or it can be used by third-party software applications with the purpose of further analysis and evaluation or extending the research by conceptualizing any new framework.

Therefore, another value for future work would be considered adding of “data prediction” feature. This feature would be possible and could be easily integrated using Artificial Intelligent (AI). This feature could be able to predict how these countries (OGD national portals) are going to publish in the coming months or a specific period. For instance, if there will be used a simple method i.e. 80/20 that means using 80% of training data and 20% of testing data, it would be easier to forecast the profiles of countries, the number of dataset publications by each publisher, types of dataset file formats and the number of file formats, publishing frequency, etc.

All these predictive data could be forecast for a specific period i.e. 6 to 12 months or probably in next 2 years.

VIII. Conclusion

Based on the study and analysis of existing frameworks for the evaluation of OGD portals, this research employed a new approach that is conceptualized and implemented through a flexible framework. This framework is considered flexible because of its adoption to any OGD portal and the ability to be available to the wider community for further research and analysis. Since the framework is composed of several components, it employes qualitative and quantitative approaches that are combined and interacted to provide compressive evaluation and monitoring results of OGD national portals.

Acknowledgment

We would like to thank the Laboratory of FINKI (Faculty of Computer Sciences and Engineering at Ss. Cyril and Methodius University) for providing all hardware and software resources for building and publication of the whole project including all components as well as hosing the project for further analysis and research. Without this reflection, would not be possible to evaluate the OGD Portals.

References

[1] Janssen, K.. “The influence of the PSI directive on open government data: An overview of recent developments. Government Information Quarterly”, 28(4), 446-456.D.E. Perry, A.L. Wolf, Foundations for the study of software architecture, ACM SIGSOFT Softw. Eng. Notes 17 (4) (1992) 40–52.

[2] European Commission. Proposal for a Directive of the European Parliament and of the Council on the re-use and commercial exploitation of public sector documents, COM (2002) 207 final 18 European Commission (2008).

[3] Vickery, G. Review of recent studies on PSI re-use and related market developments. Information Economics, Paris , 2011.

[4] Rosacker, Kirsten M., and David L. Olson. "Public sector information system critical success factors." Transforming Government: People, Process and Policy (2008).

[5] Ubaldi, B. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. Tech. rep., OECD Publishing. (2013).

[6] O. Whitehouse. Transparency and Open Government. [Online]. Available: https://obamawhitehouse.archives.gov/the-press. 2009.

[7] Magalhaes, Gustavo, Catarina Roseira, and Sharon Strover. "Open government data intermediaries: A terminology framework." Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance. 2013.

[8] Berners-Lee, T. Linked data-design issues. Tech. rep., W3C, http://www.w3.org/DesignIssues/LinkedData.html. (2006)

[9] Wang, R. & Strong, D. Beyond accuracy: What data quality means to data consumers. J. Manage. Inform. Syst. 12, 4. (1996).

[10] Wand, Y, & Wang, R. Anchoring data quality dimensions in ontological foundations. Comm. ACM 39, 11. (1996).

[11] Thorsby, J., Stowers, G. N., Wolslegel, K., & Tumbuan, E. Understanding the content and features of open data portals in American cities. Government Information Quarterly, 34(1), 53-61. (2017).

[12] Van der Waal, S., Węcel, K., Ermilov, I., Janev, V., Milošević, U., & Wainwright, M. Lifting open data portals to the data web. In Linked Open Data--Creating Knowledge Out of Interlinked Data (pp. 175-195). Springer, Cham. (2014).

[13] Umbrich, J., Neumaier, S., & Polleres, A. Quality assessment and evolution of open data portals. In 2015 3rd international conference on future internet of things and cloud (pp. 404-411). IEEE. (August, 2015).

[14] Ham, J., Koo, Y., & Lee, J. N. Provision and usage of open government data: strategic transformation paths. Industrial Management & Data Systems. (2019).

[15] Kalampokis, E., Karamanou, A., Nikolov, A., Haase, P., Cyganiak, R., Roberts, B., ... & Tarabanis, K. A. Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services. In SemStats@ ISWC. (October, 2014).

[16] Yang, S. Quality Diagnosis of Library-Related Open Government Data: Focused on Book Details API of Data for Library. Journal of the Korean Society for Information Management, 2020 37(4), 181-206.

[17] Thorsby, J., Stowers, G. N., Wolslegel, K., & Tumbuan, E. Understanding the content and features of open data portals in American cities. Government Information Quarterly, 34(1), 53-61. (2017)

[18] ODI- Global Open Data Index (Methodology - Global Open Data Index (okfn.org))

[19] Open Data Inventory—Global Index of Open Data - Open Data Inventory (opendatawatch.com), (2021).

[20] Methodology | Open Data Barometer (opendatabarometer.org), (2021).

[21] Sayogo, D. S., & Pardo, T. A. Exploring the motive for data publication in open data initiative: Linking intention to action. In 2012 45th Hawaii International Conference on System Sciences (pp. 2623-2632). IEEE.(January, 2020).

[22] Strong, D. M., Lee, Y. W., & Wang, R. Y. Data quality in context. Communications of the ACM, 40(5), 103-110. (1997).

[23] Milani, M., Bertossi, L., & Ariyan, S. Extending contexts with ontologies for multidimensional data quality assessment. In 2014 IEEE 30th International Conference on Data Engineering Workshops (pp. 242-247). IEEE. (March, 2014).

[24] Maurino A., Spahiu B., Batini C., & Viscusi G. Compliance with Open Government Data Policies: an empirical evaluation of Italian local public administrations, Twenty Second European Conference on Information Systems, Tel Aviv,(2014).

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

[26] Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly, 33(2), 325-337. (2016).

[27] Parycek, P., Höchtl, J., & Ginner, M. Open government data implementation evaluation. Journal of theoretical and applied electronic commerce research, 9(2), 80-99. (2014).

[28] Nikiforova, A., & McBride, K. Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics, 2021 58, 101539. (2021).

[29] CKAN, (https://ckan.org/about/).

[30] DKAN Open Data Platform (https://getdkan.org).

[31] Open data Montenegro (https://data.gov.me).

[32] Bosna and Herzegovina (http://opendata.ba).

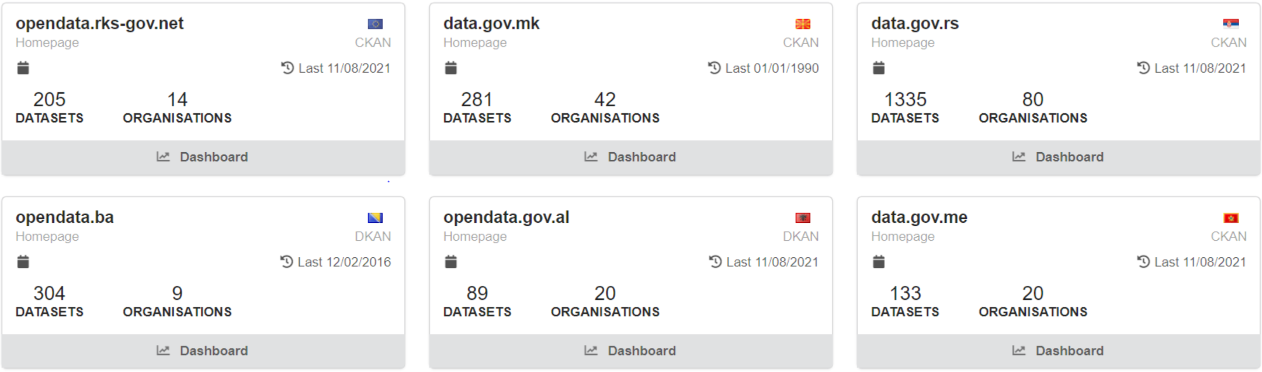
[33] North Macedonia (https://data.gov.mk/).

[34] Serbia, Data.gov.rs. (https://data.gov.rs).

[35] Albania, OpenData Faqja Kryesore (https://opendata.gov.al).

[36] Kosovo, RKS Open Data (https://opendata.rks-gov.net).

APPENDIX



opendata.rks-gov.net — Homepage — CKAN — Last 11/08/2021
205 DATASETS | 14 ORGANISATIONS
Dashboard

data.gov.mk — Homepage — CKAN — Last 01/01/1990
281 DATASETS | 42 ORGANISATIONS
Dashboard

data.gov.rs — Homepage — CKAN — Last 11/08/2021
1335 DATASETS | 80 ORGANISATIONS
Dashboard

opendata.ba — Homepage — DKAN — Last 12/02/2016
304 DATASETS | 9 ORGANISATIONS
Dashboard

opendata.gov.al — Homepage — DKAN — Last 11/08/2021
89 DATASETS | 20 ORGANISATIONS
Dashboard

data.gov.me — Homepage — CKAN — Last 11/08/2021
133 DATASETS | 20 ORGANISATIONS
Dashboard

Fig. 10. Monitoring of OGD National Portals (Front-end of Portal Developed).

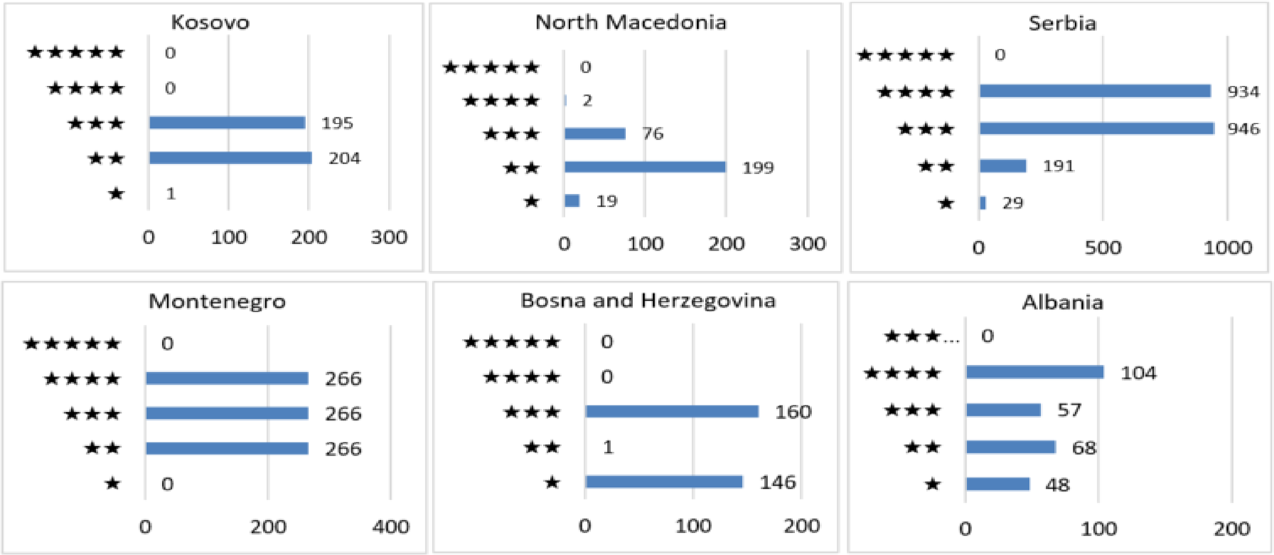


Fig. 11. Openness Evaluation (Evaluation results Presented by Graphs for each Country OGD Portal).