DEVELOPMENT FRAMEWORK FOR MONITORING AND EVALUATION OF OPEN GOVERNMENT DATA QUALITY

Vigan Raca[1], Goran Velinov[1], Betim Cico[2], Margita Kon-Popovska[1]*

[1] St Cyril and Methodius University in Skopje, Computer Science and Engineering
[2] Metropolitan University of Tirana, Computer Science and IT
[1]North Macedonia, [2]Albania

*Corresponding Author, e-mail: viganraca@gmail.com

Abstract: Open Government Data remains a worldwide challenge, in both developed and developing countries. There are various models for measuring the quality of Open Government Data but in our research paper, we have designed a framework for evaluating and Open Government Data quality. Target of evaluation are six Western Balkan countries, respectively Open Government Data portals with the possibility of extension to other countries. For assessment purposes, we have designed an application that uses combination of metrics based on two data quality dimensions, while in the background it will perform various calculations for showing results in a dashboard. The results will show the performance of each country as well as their quality progress in monthly basis.

Key words: Open Data, Quality, Dataset, Government, Portal, Evaluation.

1. INTRODUCTION

The Open Government Data (OGD) as concept has been present for more than thirteen years, though there remain differences in terminology depending on the literature and usage. Since, the concept refers to Government; it also refers to public data. In the last two decades, great efforts have been made in promoting OGD with the sole purpose of increasing transparency and public sector accountability [1]. This has also strengthened the determination of governments and civil society in trying to increase the openness of government data as much as possible.

Being open means that your government data can be used, reused, and distributed from anyone without any limitation [2]. This is due to the fact that they are open and available to anyone without restriction. Therefore, considering this fact, the quality of open government data remains a challenge that consistently harasses scientists despite numerus efforts by developing various mechanisms to improve the

quality of open data [3]. This is also considered very important because increasing the quality of data, increases the effectiveness of the use of this data [4]. Moreover, our work focuses on framework development that aims to monitor and evaluate the quality of open government data. The research cover six Western Balkans (WB) countries including Albania, Bosnia and Hercegovina, North Macedonia, Montenegro, Kosovo and Serbia, respectively OGD portals, highlighting the issues of data quality, as well as the number of involved organizations and datasets published for each country. The paper also provides the average data quality for each country and provides the opportunity for comparison noting the differences between countries.

For this purpose, we have developed a web-service that monitors six Western Balkan countries. In addition to monitoring, the web-service will have another role to collect and store the information about datasets, resources, organizations, file types of datasets and other relevant information explained in the following sections. In addition, an integral part of our work is development of a dashboard for displaying the collected information such number of published datasets, organizations, licenses, and types of datasets. Apart of this function, it also performs another function. This function is the evaluation of dataset quality from collected data performed by the web-service. Moreover, there will be able to monitor and show quality of datasets for each WB country in time. For this purpose, our designed portal in background performs various mathematical operations for extracting the data quality for each country. At the same time, it provides historical data of performed evaluations for each country by showing the progress/regress achieved and portrays it in a visualized manner.

In the first part of this paper, we discuss the data quality state of play in the Western Balkans; in the second part addresses the issues of setting up the web-service for gathering target information; the fourth part discusses the preparation of data for evaluation as well in order to measure data quality and the last part presents results shown in portal developed for this purpose.

2. STATE OF THE ART OF OGD IN BALKAN REGION

Given the fact that the countries of the Western Balkans are involved later in the open data initiative compared to developed EU countries, it will take time for these countries to develop e-governance and provide qualitative. A successful implementation of e-governance will certainly improve public services and will help strengthening of society [9]. In this context, despite the fact that most WB countries have already become part of the Open Government Partnership excluding Kosovo for political reasons, a lot remains to be done towards improving data quality [10]. OGP is the main promoter in encouraging the opening of government data by expanding the range of member institutions and setting technical criteria based on "action plans" to improve the data quality to use them as efficiently as possible. This partnership also pushed WB countries to develop their national portals dedicated to

the publication of open public data [7, 8]. Since these countries have developed their own OGD portals, the structure of these portals differs from the other. A simple comparison of all these portals reveals that all of them have in place well-organized resource including available datasets, dataset file formats and the respective licenses, organizations joined and provides possibility of using an Application Programmable Interface (API) [13].

When it comes to APIs, it is important to note that each OGD portal in each country uses an open-source API that is dedicated to publishing data into portal as well as it enables using of published data by third party applications. In addition, this API also improves quality of datasets during publication process, so the member organizations for each selected dataset for publication will be subject of quality check provided by this open-source API. In addition, table 1 shows the portal data including the number of datasets, organizations and the type of APIs used.

Table 1. Open Government Data Portals Publication

Country	***Datasets***	***Organisations***	***API***
Albania	*88*	*20*	*CKAN*
Bosna & Herzegovina	*305*	*10*	*DKAN*
Kosovo	*205*	*14*	*CKAN*
North Macedonia	*266*	*41*	*CKAN*
Montenegro	*123*	*19*	*CKAN**
Serbia	*303*	*56*	*CKAN**

It is also important to note that the high number of datasets does not reflect quality of dataset. As for APIs, these countries have mainly employed two types of APIs, CKAN and DKAN. Serbia and Montenegro have developed their API but it is based on CKAN so for those countries we have used following acronym (CKAN *).

CKAN is an open-source data platform developed by the Open Knowledge Foundation, a non-profit organization that aims to promote the openness of all forms of knowledge. In other words, CKAN is a tool for making open data portals. It helps to manage and publish data collection [14].

DKAN is an open data platform, based on the CKAN, with a full suite of cataloging, publishing and visualization features that allows governments, nonprofits and universities to easily publish data to the public [15].

Regarding the data quality, the WB countries do not apply any strict technical rule such as publication format which aims to regulate the quality of data.. It has regulated this, so each format is published in all formats accepted by the 5 Star scheme [5]. In this regard, there is not defined any standards excluding Montenegro. Moreover, they neglect the rules set by the "Open Knowledge Foundation" on formats (in one of the open formats), the definition of licenses that support data opening (openness) as well as the published data to be discoverable [6].

Therefore, in the framework of this research, we have designed a tool that will monitor resources publication in all six OGD portals and will enable measuring of

data quality showing in a dashboard designed for this purpose. This application tool is authentic and does not exist any other tool which perform evaluation and monitoring of WB countries. Globally in worldwide exist different monitoring portals which perform evaluation and monitoring of OGD such (global open data index, open data barometer) [11, 12], but ours differs from those because the framework is designed to be fit with existing situation of OGD portals in WB also the algorithm used is different with other barometers used before.

3. METHODOLOGY

For monitoring and measuring quality of datasets published in OGD portals, initially this paper builds a web-service for this purpose as well as a portal for showing monitored resources and evaluation results. These two components consist other processes but we have grouped in two main phases.

Preparation phase mainly covers the web-service part. It has two key specific objectives:

(1) monitoring of six WB portals by using the APIs made available for this purpose.
(2) collecting and storing information in our database such name of joined organizations, file format types, the number of datasets, names of datasets available, licenses and the publication or last update date.

This web service continuously run and perform the mentioned process, it also ensures updates of our database if there is any new information updated in one of the six OGD portals which are target of monitoring.

In addition, part of this phase is data preparation that means a process of preparing data for further evaluation. In this regard, we have designed two separate processes for this purpose including (1) the data cleaning process that removes unnecessary data from the database and (2) a validation process, where it is much more important and ensures that processed data to be valid and ready for Phase II of the evaluation. Validation process consist regulating of data types, date & time formats, Cyrillic to Latin alphabet recognition since here we have to deal with 6 different languages of the OGD portals and the native language of each portal is used as default. In this respect, the data collected and inserted into our database have used the native language of portal. It argues that validation process is more than necessary.

Evaluation phase that covers the designing of an algorithm for evaluation of data quality. For this purpose, we have designed to models for data quality evaluation:

(1) traditional data quality model which uses four metrics like (accessibility, availability, discoverability and timeless). This model derives from traditional DQ six dimensions but we have modified and made it compatible for our datasets and
(2) 5- Star schema model that derives from Berners-Lee for linked open data but modified version and made compatible with file formats of WB OGD

portals. This model comparing with previous one uses file format of dataset as metric for evaluation.

Moreover, within this phase various calculation are performed with the aim of effective production of the data quality results for each country shown in our designed WB data quality portal.

4. DATA COLLECTION AND DATA PREPARATION

As defined in the methodology, data preparation is the main purpose of the first phase. According to the fact that we still do not have any data in our database for monitored OGD portals that we targeted in this paper, this makes further analysis impossible. So, as an initial step is the development of a web service which will have two main objectives: monitoring of OGD portals and storing this monitored information into our database. In figure 1 we have shown the block schema for stages from collection of data to preparation for further analysis. In table 2 we have presented what data the web-service will collect.

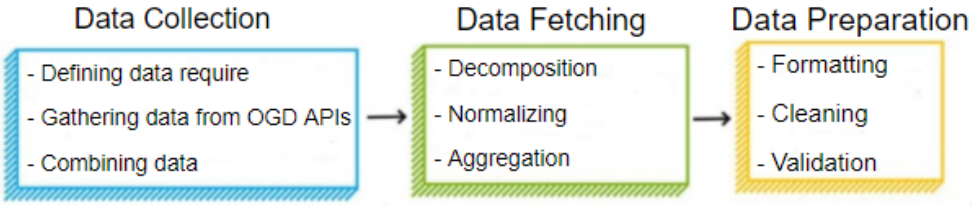

Fig. 1. Data preparation stages

Table 2. Target data to be collected using designed Web-Service

Target Data	***Information Type***	***Get from***
Dataset's Name	*String*	*API*
Organization's name	*String*	*API*
Dataset File Format	*String*	*Web*
License Types	*String*	*Web*
Published Date	*Date & Time*	*API*

4.1. Data Collection Process

Data preparation as process means the collection of this data as a pre-qualifying step for further analysis. In this respect, we have developed a web-service with the aim to monitor all six OGD portals, with special emphasis on the information described in table 2. This web-service is developed in Microsoft .NET platform, based on MVC technology. The challenges that arose during the development of the web-service were on establishing of communication with APIs of OGD portals.

Consider the fact that each portal target of our research has its own API that differ from each other. This complicates process of interconnection of our web-service with rest APIs because there is not any model applicable that will be compatible to all OG portals, but it has been adapted individually for each of them.

In addition, the web service will have ability to be run continuously to monitor in real-time the above-mentioned portals and to collect target information. It also can be configured to be run on schedule in specific time. Figure 2 shows the scheme of operation of the web service.

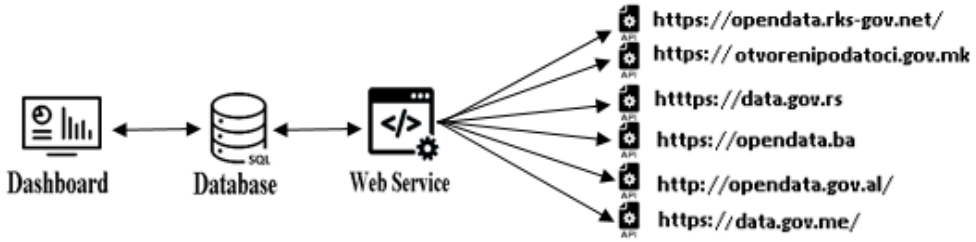

Fig. 2. Communication of Web Service with OGD Portal APIs [16-21]

s shown in the figure above, the web service besides the monitoring role, it covers the other role of inserting monitored data into database for further analysis. In the below box, we have presented fragments of the web-service code that performs collection of target information defined.

```
driver.Manage().Window.Maximize();
        driver.Navigate().GoToUrl("http://opendata.gov.al/api/sq/institutions");
        string content = driver.FindElement(By.TagName("body")).Text;
        JArray json = JArray.Parse(content);

        foreach (var item in json.Children())
        {
           Organisations objOrg = new Organisations();
           objOrg.PortalID = PortalID;
           objOrg.Organisation = item["name"].ToString();
           objOrg.OrganisationURL = item["id"].ToString();
           string result = GetDatasetPost(item["id"].ToString()).Result;
           JArray json2 = JArray.Parse(result);
```

Fig. 3. Getting dataset information from portal (Albania APIs)

4.2. Data Fetching Process

Once the web service has managed to provide and collects the defined information, there is necessary to store this information somewhere to further processing and analysis. Given that, we have developed a modest small database will be available for built-in web service. At the same time, when new information is available in one of the six portals the web-service will have ability to detect and store the updates in database.

In addition, the data insertion process is based on algorithm designed with the aim of ensuring that all data collected will successfully inserted in our database. After each process will be performed a snapshot, it will ensure if anything goes wrong during bulk insert process, there is no need to reinitialize.

Algorithm 1 Inserting data from six OGD Portals

```
define fetch( OGD portals):
// bulk insert dataset names
snapshot = sysdate()
portal info:
//get the list of datasets available in portal
datasets_list = get_datasets(portal .api)
for id in datasets_list :
dataset = get(portal .api , id)
// bulk insert datasets into database
store_in_db(dataset , snapshot)
//get the list of organizations
resource_list = get_resources(portal .api)
for id in resource_list :
resource = get(portal .api , id)
// bulk insert organizations into database
store_in_db(resource , snapshot)
sleep()
```

The various problems and challenges have been faced during data collection and insertion process. Starting to "date & time" data types, which depends on countries language collisions used especially for Serbia and Macedonia because they have used Cyrillic Alphabet. For this purpose, we have used "Convert" function to adapt to a proper format. In the figure 4 is shown designed relational database diagram.

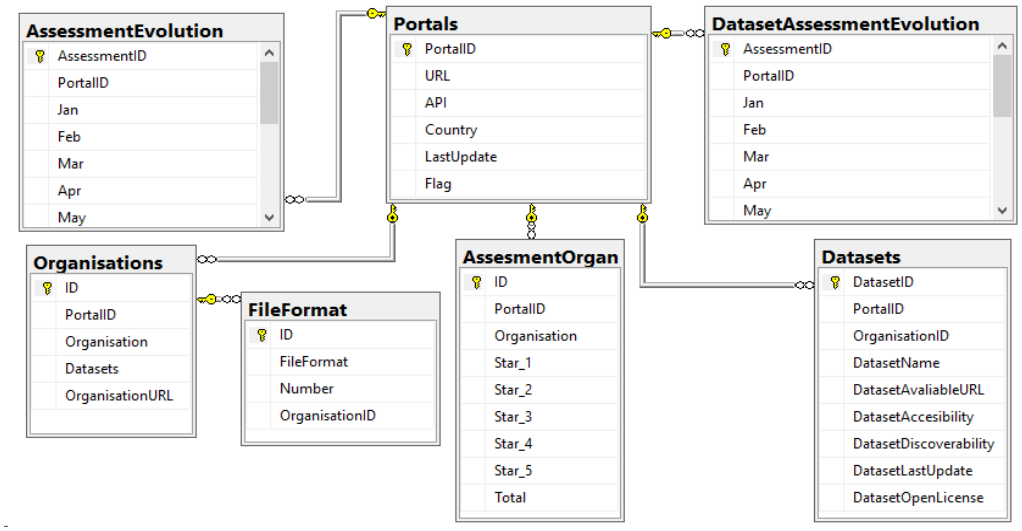

Fig. 4. ER Model for database diagram used

4.3. Data Preparation Process

Completing of successfully data import process for all six OGD portals, will raise the need for preparing data as final step for further analysis and evaluation. In this respect, we will use two separate processes: Data Cleaning and Data Validation.

Data Cleaning as a process aims to remove unnecessary data which does not have any impact on our results. Thus, for this purpose, initially after new information is uploaded in our database, a special stored procedure cleans the unnecessary data. In the context of data cleaning, it was necessary to remove following information:

- incorrect names of organizations,
- incorrect names of datasets,
- wrong file formats,
- null values and unknown values

In figure 5 we have shown T-SQL stored procedure prepared for performing data validation and data cleaning on collected data.

```
--Data Cleaning Process–
UPDATE dbo.Datasets
      set DatasetOpenLicense =
      case DatasetOpenLicense
      when 'Licenca nuk është specifikuar' then 'Undefined'
      when N'Лиценца није специфицирана' then 'Undefined'
      when 'Other (Public Domain)' then 'Other Public'
      when 'Other (Open)' then 'Other Public'
      else DatasetOpenLicense
      END
GO
      UPDATE dbo.Datasets
      SET DatasetOpenLicense = 'Undefined' where (DatasetOpenLicense
like '%1%' or DatasetOpenLicense like '%2%' or DatasetOpenLicense like
'%3%' or DatasetOpenLicense like '%4%'…..
      UPDATE dbo.Datasets
      set DatasetLastUpdate = null
      where DatasetLastUpdate = '1900-01-01 00:00:00.000'

--Data Validation Process--
      UPDATE dbo.FileFormat
      SET   FileFormat =
            case  FileFormat
            when '-godinu.xlsx' then 'XLSX'
            when 'geojson' then 'JSON'
            when 'reojсон' then 'JSON'
            when 'gradiste-budzet-.xlsx' then 'XLSX'
            when '.pdf' then 'PDF'
            when 'XLSC' then 'XLSX'
      ELSE FileFormat . . .
      END
```

Fig. 5. Procedure for data cleaning and data validation

Data Validation - it is important that once the data cleaning process is complete, this data should be subjected to another validation process. Since there exist various validation types but based on our collected data we have used following types:

- Datatype Validation (applying unique datatype format since there exist six different data formats depends on portal and native language used)
- Consistence Expressions (it was necessary due to multiple expressions of licensed used by portals) and
- No Null values (eliminating empty strings and replacing null values)

This is very important especially in measuring the quality of datasets as portals have most of their datasets published using inadequate extensions such as in some countries XLS in some XSL or JSON or GEJSON or in some cases have published using the alphabet Cyrillic in the extensions they have published. In addition, validation has been applied in the format (Date & Time) as it has been necessary due to inadequate use of published formats. For example, some have used the format dd-mm-YYYY, while others have used the format of dd / mm / YYYY or mm / dd / YYYY, thereby validation has been more than necessary. In figure 4 we have presented code samples used in SQL (stored procedure) which performs the two processes mentioned above.

5. EVALUATION MODEL

Since, now we have ensured all necessary data through the web-service and after they have undergone the repurchase process, the preconditions for quality evaluation have been prepared based on the criteria mentioned in the methodology. In this context, a general evaluation cannot be done as we are evaluating two dimension including dataset quality and openness.

5.1. Data Quality Evaluation

To evaluate the dataset quality, we used a combination of DQ metrics by selecting only four metrics but based on six dimensional indicators. This is due to compatibility of four metrics to all six WB OGD countries portal. In table 3 we have shown description of each metric used.

Table 3. Data Quality Dimension

Metric	Description
Availability	*If Dataset is publicity available*
Accessibility	*If Dataset is accessible for download*
Discoverability	*If the data inside the dataset are searchable*
Timeless	*If Dataset is up-to date regularly in time*

There is important to note that all metrics except for “timeless” have been given the status defined during the data import as 0 or 1. While “timeless” metric we have

used different approach. In following figure 6 we have shown the calculation performed.

```
SELECT CONVERT(numeric(10,2), DATEDIFF(day,
DatasetLastUpdate,getdate())) as Diff *into #table
SET @t1 = (SELECT ISNULL(((365-AVG(Diff))/365),0) FROM #table WHERE
Diff <=365 );
. . .
```

Fig. 6. Timeless metric calculation

Initially, we have used the earliest and last date of publication. Then the calculation is done by dividing into slots for five years. This is because analysed portals content datasets within five years. Each slot is named for ex. "t1" derives the average time difference of publication within a year (365). It continues to increase for 1 per year and we have situations of using 1825 days. Moreover, table 4 summarizes all the results produced by four metrics for each country:

Table 4. Dataset Dimension Results

Country	**Avail.**	**Access.**	**Discover.**	**Timeless**
Albania	*1*	*1*	*0.49*	*0.44*
Bosna & Herzegovina	*1*	*1*	*0*	*0.03*
Kosovo	*1*	*1*	*1*	*0.18*
Montenegro	*1*	*1*	*1*	*0.37*
North Macedonia	*1*	*1*	*0.81*	*0.36*
Serbia	*1*	*1*	*0*	*0.29*

5.2. File Format Openness Evaluation

The “Openness Dimension” is another calculation performed which is based on published file formats. Referring to table 5, it shows the "metrics" that have been used, but here the metrics are marked with "star" and it means that higher star rates, the format is more qualitative.

Table 5. Data Openness Dimension

Rate	***Description***
★	*Available on the web in PDF or Word unstructured format*
★★	*Available as machine-readable structured data*
★★★	*Available in non-proprietary format (e.g. CSV instead of excel)*
★★★★	*Available in open standards from w3c (e.g. Html, Xml, Json)*
★★★★★	*Available in above plus link data to other data sources*

In addition, table 6 shows how the datasets are grouped based on file format extensions for each country.

Table 6. Data Openness Dimension Results

Country	★	★★	★★★	★★★★	★★★★★
Datatype Formats	*Pdf, Doc*	*Xls, Xlsx*	*Csv*	*Html, Xml, Json*	*Linked*
Kosovo	*1*	*204*	*195*	*0*	*0*
North Macedonia	*19*	*191*	*62*	*2*	*0*
Serbia	*9*	*93*	*93*	*59*	*0*
Bosna & Herzegovina	*147*	*0*	*160*	*0*	*0*
Albania	*47*	*65*	*57*	*104*	*0*
Montenegro	*0*	*246*	*123*	*246*	*0*

6. RESULTS AND COMPARISON

Since the main purpose of this paper is the development of an "entity framework", which is implemented through an application, using an integrated dashboard for displaying produced results. However, there will be performed two separate calculations for each measurement. To derive the required average of the results produced by metrics of "Dataset Dimension" framework we have calculated based on formula (1).

$$\frac{\sum(Avalib.) + \sum(Access.) + \sum(Discov.) + \sum(Timelss)}{4} \quad (1)$$

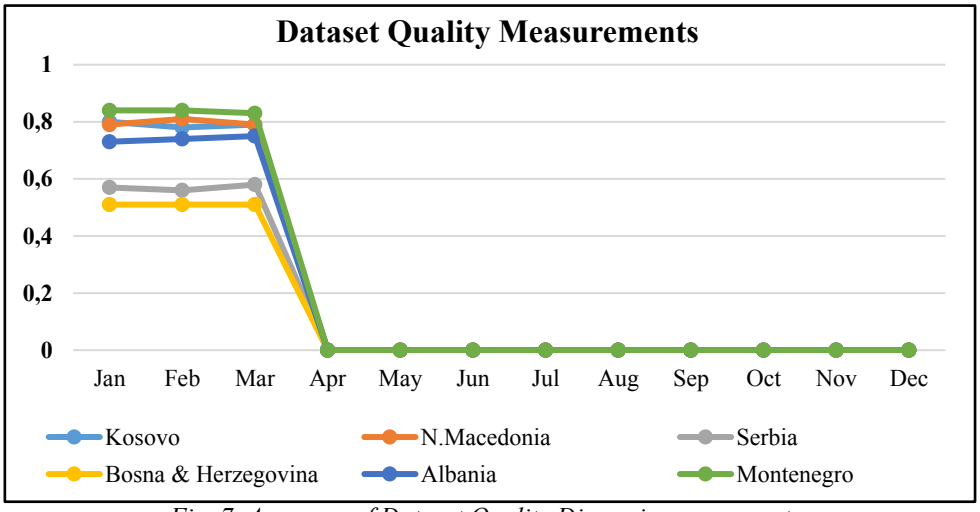

Fig. 7. Averages of Dataset Quality Dimensions per country

The same goes for the other metrics of "Openness Dimension", so our dashboard presents the results, but here we have planned another formula, which adds the

average of all "rate stars" by proportioning them to the total number of datasets per country. For calculating this, we have used formula (2).

$$\frac{\sum(1\ star)*1 + \sum(2\ star)*2+\sum(3\ star)*3+\sum(4\ star)*4+\sum(5\ star)*5}{\sum total\ datasets} \quad (2)$$

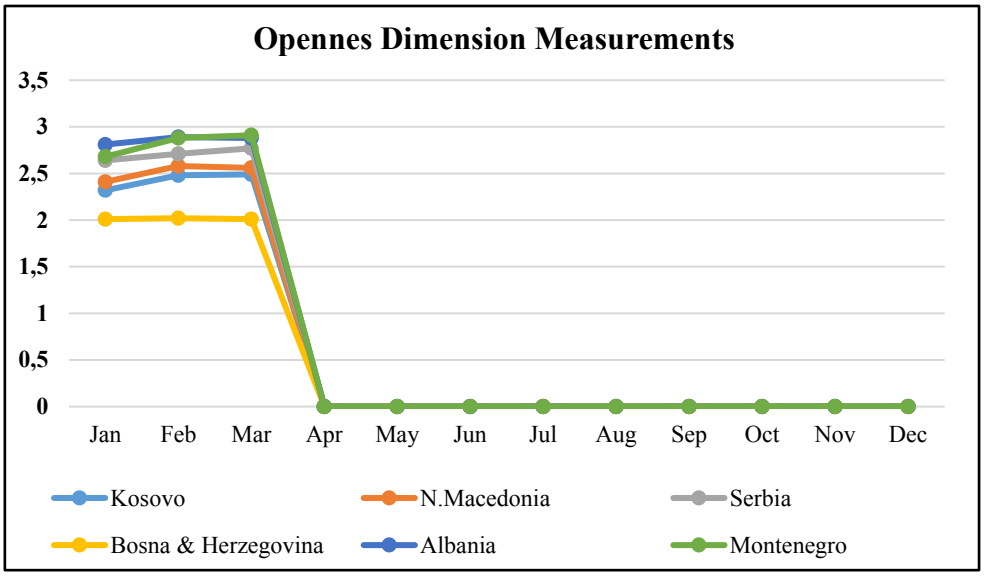

Fig. 8. Averages of Openness Quality per country

7. CONCLUSION

This research initially addresses the situation of OGD state of play in Western Balkan in the context of measuring and evaluating of published datasets into their national OGD portals using an designed application for performing this evaluation

Used methodology have produced effective results, while combination of Openness Dimension and Dataset Dimension ensured productive results.

Initially, a web service prototype is designed to collect dataset information as precondition for further analysis and evaluation. A dashboard is developed for monitoring and presenting evaluation results for six WB countries included in this research.

It important to note that our designed prototype absolutely can be extended to other counties in a simple way. Only the API of target OGD Portal should be integrated, while the application will have affinity to show results for every new portal connected. Since our research addressed the development of a framework for monitoring and evaluation OGD quality, discussions of countries performance was not our target, but the framework designed and its implementation, which is authentic and has not yet been used by anyone else in Balkan and wide Region.

REFERENCES

[1] Jetzek, T., Avital, M., Bjorn-Andersen, N. Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9, No. 2, 2014, pp. 100-120.

[2] Bauer, F., Kaltenböck, M. *Linked open data: The essentials*. Edition mono/monochrom, Vienna, 2011 (710 p.).

[3] Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Sheets, D. Tabulator: Exploring and analyzing linked data on the semantic web. *Proceedings of the 3rd International Semantic Web User Interaction Workshop,* 2006, p. 159.

[4] Janssen, M., Charalabidis, Y., Zuiderwijk, A. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Vol. 29, No. 4, 2012, pp. 258-268.

[5] 5 Star Schema (www.5stardata.info).

[6] Open Knowledge Foundation, Open data handbook, (www.opendatahandbook.org/guide/en/what-is-open-data/).

[7] Raca, V., Veljković, N., Velinov, G., Stoimenov, L., Kon-Popovska, M. Real-time monitoring and assessing open government data: A case study of the Western Balkan countries. *International Conference on ICT Innovations,* Springer, Cham, Skopje, 2020, pp. 189-201.

[8] Raca, V., Veljković, N., Velinov, G. Open government data quality assessment in Western Balkan countries. *18th International Conference e-Society 2020*, IADIS. Sofia, 2020, pp. 18-26.

[9] Rahmani, B., Idrizi, F., Halili, F. An Analysis of development of e-government in Macedonia and creating a mobile application for the secretariat for European affairs. *International Journal on IT and Security*, Vol. 8, No. 2, 2016, pp. 15-22.

[10] Open Government Partnership (www.opengovpartnership.org).

[11] Global Open Data Index (www.okfn.org).

[12] Open Data Barometer (www.opendatabarometer.org).

[13] Application Program Interface (www.webopedia.com/TERM/A).

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

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

[16] Open data Montenegro (www.data.gov.me).

[17] Bosna and Herzegovina (www.opendata.ba).

[18] North Macedonia (www.otvorenipodatoci.gov.mk/).

[19] Serbia, Data.gov.rs. (www.data.gov.rs).

[20] Albania, OpenData Faqja Kryesore (www.opendata.gov.al).

[21] Kosovo, RKS Open Data (www.opendata.rks-gov.net).

Information about the authors:

Vigan Raca – PhD candidate at St Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering. Interested areas of scientific research are Data Quality, Open Data, Data Warehouse, Database Optimization etc.

Goran Velinov – Associated Professor at St Cyril and Methodius University, Faculty of Computer Science and Engineering. Research interests in fields of Databases, Big Data, Blockchain Database, Datawarehouse and Database Optimization.

Betim Cico – Professor at Metropolitan University of Tirana, research interest in fields of Data Mining, High Performance Computing, Data Analysis and Visualization.

Margita Kon-Popovska – Professor at St Cyril and Methodius University, Faculty of Computer Science and Engineering. Research interest in fields of Data Mining, Data Analysis, Big Data, Data Quality, Database Optimization.

Manuscript received on 31 March 2021
Renewed version received on 22 April 2021