# Towards Open Government Data Quality Improvements through an Assessment Framework

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

[1]Ss Cyril and Methodius University, Faculty of CSE, Skopje, North Macedonia
[2]Epoka University, Department of Computer Engineering, Tirana, Albania

**Abstract.** Open Government Data (OGD) has recently gained attention worldwide and aims to promote accountability and transparency of the governments. At the same time, the quality of OGD remains an important factor for effective use. Different frameworks have been proposed, a few of them integrated at the application level, but our proposed framework uses a different approach. In addition to the OGD quality assessment, the proposed framework will be able to generate compressive and comparative results aimed at providing some recommendations for improvement of the quality of data of public sector bodies as main data producers. This approach has been applied to Western Balkans OGD national portals but can be expanded to other countries.

**Keywords:** Open Data, Quality, Public Sector, Assessment, Metric

## 1 Introduction

The notion of open government data (OGD) has been present over the last decade, and scientific interest in and the attention given to this subject have consistently increased through the years. The main motive comes from the intention to promote transparency and accountability in supporting e-Government, but it has recently undergone a change in bringing forward the concept of open governments [1]. This paper has gathered insights regarding both of the above areas discussed, but more attention has been given to identifying OGD quality issues. The purpose of this research is evaluation of OGD quality to identify data quality issues of public sector bodies. The extended version of the research work has been presented [2],[3] while in this paper, we will try to conclude the results and show how the results have influenced public sector bodies of Western Balkans OGD portals to improve their quality of data due to the assessment results. Moreover, the paper is organised into the following sections: the first chapter provides a short introduction as preliminaries for OGD and the role of data quality for effective data usage. The second presents the methodology used for the frameworks conceptualisation. The third is the analysis of the current situation with the OGD portals (Western Balkans OGD portals)

* Corresponding author: viganraca@gmail.com

challenges and weaknesses. Implementation of the framework and, finally, the assessment results.

## 2 Research methodology

The methodology used in this paper is based on an analysis of several case studies and frameworks used for evaluating OGD [4][5]. In comparison to others, in this paper, we have conceptualised a framework composed of several components, such as open data analysis, collection, data preparation and validation, data evaluation, and, finally, the results, leaving the possibility for the follow-up to other researchers who might have the interest in deeper research. That is possible as the data were collected locally and not directly evaluated at the portal level. Qualitative and quantitative methods have been combined to produce a framework model [6]. Table 1 shows the combination of both methods for the framework conceptualisation.

**Table 1.** Quantitative and Qualitative Model used

| | 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 | Publishers' Names |
| | Groups | Public Sector Bodies |
| | Licenses | Types of Licenses |

## 3 Analysis

Since the main objective of this research is to build a framework, some prerequisites have been identified that will support its conceptualisation, and the definition of proper metrics. Generally, the OGD portals provide vast amounts of information that are not all relevant, so setting some criteria for selecting the relevant information is considered the most valuable part of the analysis. The analysis points out huge differences between OGD portals, both visually and in terms of content. Regarding the content of resources, there is a significant difference between published resources such as datasets, public sector bodies, dataset formats, the language used, types, groups, licenses, and other organisational aspects. In Table 2, we presented a general overview of resources publication on the OGD national portals of the Western Balkans [7-12].

**Table 2.** Open Government Data Portals Resources Publication

| Country | OGD URLs | API Model | Datasets | Public Bodies |
|---|---|---|---|---|
| Albania | https://opendata.gov.al/ | CKAN | 89 | 20 |
| Bosnia and Herzegovina | https://opendata.ba | DKAN | 304 | 9 |
| Kosovo | https://opendata.rks-gov.net/ | CKAN | 205 | 14 |
| Montenegro | https://data.gov.me | CKAN* | 281 | 42 |
| North Macedonia | https://data.gov.mk/ | CKAN | 133 | 20 |
| Serbia | htttps://data.gov.rs | CKAN* | 1335 | 80 |

## 3 Assessment of OGD Quality

With reference to the identification of resources in the analysis section above, the proposed framework will have three key functions. First, the openness of governments through dataset publication format. Second, data quality through the observability of datasets based on the existence of information about the dataset published. Third, the quality of data (rows and records) within the dataset. For this purpose, we have named dimensions differently, Quantitative Metrics and Qualitative Metrics. These metrics are explained in the following sections.

### 3.2 Quantitative Metrics Calculation

The most simplified dimension based on quantitative methodology is conceptualised to monitor the OGD Portals by counting and grouping datasets based on file format extension. The purpose of these metrics is to measure the openness of governments based on the 5 star-schema model of Berners-Lee (2006). Thus, for calculating averages per country (OGD portal), a series of mathematical calculations will be performed using formula (1):

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

Equation (1) calculates the average government openness by adding the whole number of datasets rated with 1 star, then with 2 stars, and 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, some public sector bodies have published their datasets in multiple file formats (e.g. CSV and JSON). In such situations, the formula above (1) does not guarantee the accuracy of results, and we have applied the formulas (2) and (3):

$$\delta = \frac{\sum H\ (n\ star)*n}{\sum H\ Datasets} \qquad (2) \qquad f(x) = \gamma + \delta \qquad (3)$$

H – Means the highest number of stars of a dataset.

### 3.2 Qualitative Metrics Calculation

In contrast, qualitative metrics are grouped into two types: datasets metrics and data metrics. Also, the rates here differ; in both groups, the minimal value per metric is rated with “0” zero, and the maximum value is “1”. Even though there are two separate groups of metrics (availability, accessibility, discoverability, and timelessness) attributed to dataset quality and (completeness, uniqueness, consistency, and validity) attributed to data quality, the calculation will be performed in the same way for both groups of metrics. In addition, the formula (4) calculates the average at the OGD portal level, summing all obtained results per metric in proportion to the number of used metrics:

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

Compared to datasets metrics, here, each metric is important and involves a lot of calculations. Incomplete datasets can be "null" values or empty strings. For this purpose, we have used formula (5), while for uniqueness metric and consistency (6), (7) and validity metrics are incorporated using formula (8).

$$\varphi = \frac{\sum(\alpha) * \sum(\beta) - \sum(\times)}{\sum(\alpha) * \sum(\beta)} \quad (5) \qquad \theta = \frac{\sum(\Omega)}{\sum(\alpha)} \quad (6) \qquad \varepsilon = \frac{\sum(\mu)}{\sum(\alpha)} \quad (7)$$

α - Total number of records in dataset
β - Total number of columns
X -Total incorrect records

Ω - Total number of duplicate records
μ - Total number of inconsistent records

$$\upsilon = \frac{\sum(\omega)}{\sum(\alpha) * \sum(\kappa)} \quad (8)$$

ω - Total number of non-valid data
κ - Total number of non-valid columns

## 4 Assessment Results

The application of formulas discussed in the above section has produced varying results. Let us begin with quantitative metrics, assessing the respective openness of governments through published datasets. In Table 3, we have shown results for each country based on a 5-star schema evaluation model.

**Table 3.** Openness Metric Calculation per Country

| Country | * | ** | *** | **** | ***** |
|---|---|---|---|---|---|
| Albania | 48 | 68 | 57 | 104 | 0 |
| Bosnia and Herzegovina | 146 | 1 | 160 | 0 | 0 |
| Kosovo | 1 | 204 | 195 | 0 | 0 |
| Montenegro | 0 | 266 | 133 | 266 | 0 |
| North Macedonia | 19 | 199 | 76 | 2 | 0 |
| Serbia | 29 | 191 | 946 | 934 | 0 |

Therefore, in a visual view, Fig.1 presents the differences among countries based on the metric.

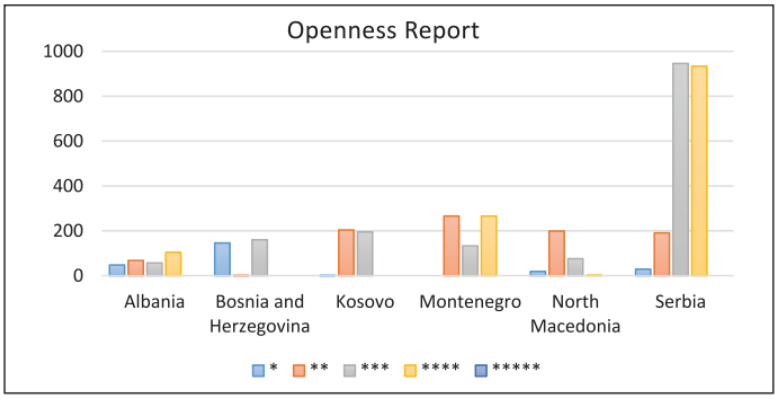


**Fig. 1.** Openness Report Generation

Regarding the Dataset Quality metrics, Table 4 shows the average results:

| Country | Availability | Accessibility | Discoverability | Timeless |
|---|---|---|---|---|
| Albania | 1 | 1 | 0.48 | 0.41 |
| Bosnia 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 |

**Table 4.** Dataset Metrics Calculation per Country

Before discussing the data quality dimension results, Table 5 presents the correct and incorrect datasets at the central and local government levels (public sector bodies).

**Table 5.** Local and Central Level Results

| Country | Datasets Evaluated | Correct Datasets | Incorrect Datasets |
|---|---|---|---|
| North Macedonia | 251 | 23 | 228 |
| Montenegro | 79 | 53 | 26 |
| Kosovo | 196 | 4 | 192 |
| Albania | 40 | 16 | 24 |
| Serbia | 331 | 95 | 236 |
| Bosnia and Herzegovina | 159 | 159 | 0 |

Regarding Data Quality Assessment results, Table 6 shows the obtained results generated by the application of formulas (5), (6), and (7).

**Table 6.** Data Quality Assessment Results

| Country | Completeness | Uniqueness | Consistency | Validity |
|---|---|---|---|---|
| Albania | 0.91 | 0.96 | 0.6 | 0.85 |
| Bosnia and Herzegovina | 0.99 | 0.98 | 1 | 0.85 |
| Montenegro | 0.99 | 0.99 | 0.33 | 0.14 |
| North Macedonia | 0.84 | 0.83 | 0.91 | 0.25 |
| Kosovo | 0.98 | 0.99 | 0.98 | 0.27 |
| Serbia | 0.94 | 0.93 | 0.71 | 0.31 |

The validity metric is considered the most problematic and complex metric because of other sub-metrics. Thus, the application of formula (8) has produced results presented in Table 7.

**Table 7.** Validity Sub-Metrics Assessment Results

| Country | Numeric | Country Code | Names/Surnames | Dates |
|---|---|---|---|---|
| Albania | 0.92 | 0.99 | 0.65 | N/A |
| Bosnia and Herzegovina | 0.02 | 0.83 | 0.44 | N/A |
| Montenegro | N/A | N/A | 0.14 | N/A |
| North Macedonia | 0.52 | 0.78 | 0.46 | 0.25 |
| Kosovo | 0.27 | N/A | 0.25 | 0.30 |
| Serbia | 0.21 | 0.54 | 0.29 | 0.16 |

Finally, in Figure 2 and Figure 3, we will visually present the comparison of assessment results for both dataset and data quality.

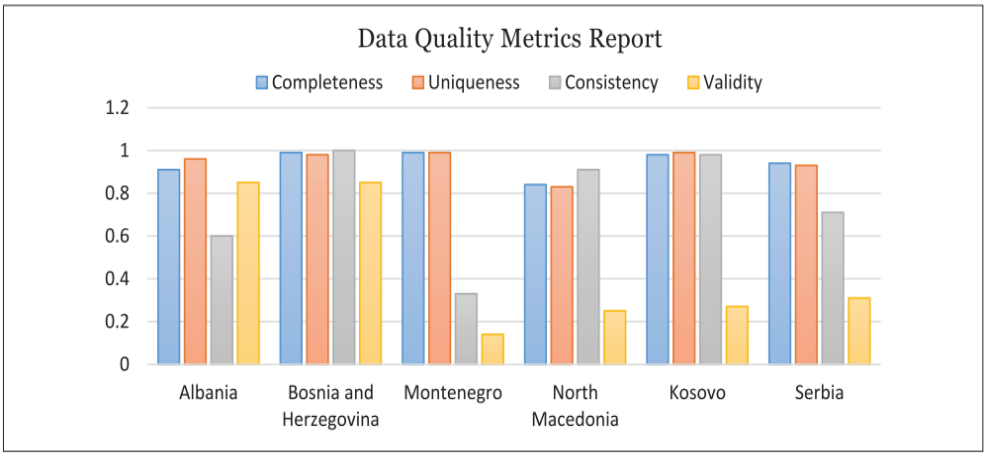


**Fig. 2.** Data Quality Report per Country

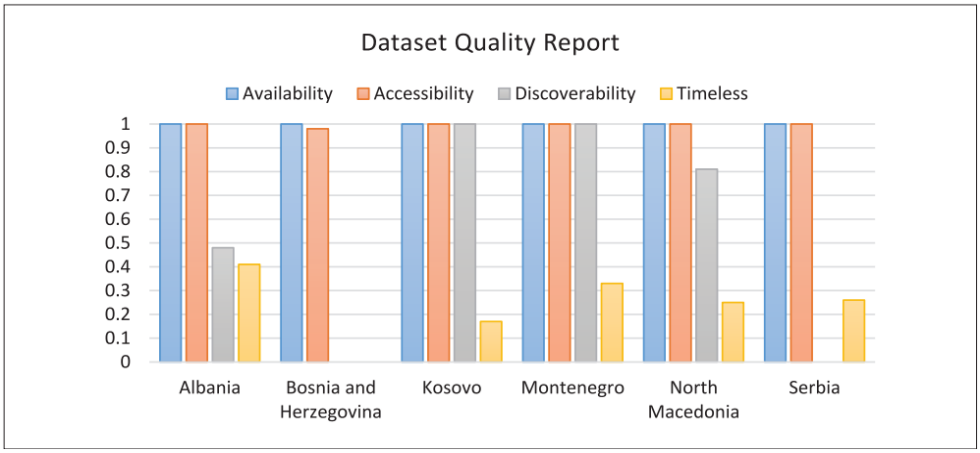


**Fig. 3.** Dataset Quality Report per Country

## Conclusion

The results presented in this research paper as a summary of the previous research work show the importance of data quality for public sector bodies. The presented results can be used to push and encourage governments and public sector bodies to increase control over their data and continuously improve the data quality. Using a comparative and comprehensive approach has strengthened the proposed framework distinguishing it from other frameworks released. Meanwhile, the adoption of the framework by other countries out of Western Balkans OGD portals would be considered a new challenge and require further research.

## References


[1] Allison, B. Open government collaboration, transparency, and participation in practice (pp. 257–265). O'Reilly Media, 2010.

[2] Raca, Vigan, et al. "A Framework for Evaluation and Improvement of Open Government Data Quality: Application to the Western Balkans National Open Data Portals." SAGE Open 12.2 (2022): 21582440221104813.

[3] Raca, Vigan, et al. "Application-based Framework for Analysis, Monitoring and Evaluation of National Open Data Portals." International Journal of Advanced Computer Science and Applications 12.11 (2021).

[4] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 16–52S. (2009).

[5] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. Quality assessment methodologies for linked open data. Semantic Web Journal, 1(5), 1–31. (2012).

[6] 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 Quarterly, 33(2), 325–337. (2016)

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

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

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

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

[11] Albania, OpenData (https://opendata.gov.al).

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

**Vigan Raća** obtained his MSc degree from South East European University in North Macedonia. Currently, he is a PhD candidate at Ss Cyril and Methodius University in Skopje, North Macedonia.