7th International Conference on
Combinatorics, Cryptography, Computer Science and Computing
November. 16-17, 2022

Understanding, Predicting and Controlling the Status of Organizations that Provide Open Government Data

Mohammad Moradi, Mojtaba Mazoochi
ICT Research Institute
Tehran, Iran
mohammad.moradi@ut.ac.ir; mazoochi@itrc.ac.ir

ABSTRACT

Open data is non-confidential data that is made available without any restrictions on use or distribution. Open government data (OGD) is a tool to empower citizens and give them access and permission to use data produced by the government organizations, so that they can use, store, redistribute and integrate the data with other data sources. Providing information in the form of open data leads to reducing corruption, gaining public trust and creating a democratic society. Also, open data provide more possibilities for monitoring governance activities. The aim of this research is to understand, predict and control the complex system of communication between citizens and organizations providing open government data in order to benefit from the advantages of open data. For this purpose, in the understanding phase, a series of comprehensive and complete evaluation criteria of organizations based on library studies have been presented. Also, the data extraction of the organizations present in the open government data portal has been done. Then, data mining techniques including the decision tree classification model has been used to predict and control the status of organizations. The accuracy of the created classification model was 90.91, which showed that the created model was desirable. By using the presented classification model, managers will be able to predict the status of organizations. Also, they can apply the necessary changes on the organizations in order to change the status of the organization to the desired status.

KEYWORDS: open government data, data mining, classification

1 INTRODUCTION

In general, all knowledge, awareness, possessions, statistics, identifiers, backgrounds and assumptions are considered data. Basically, data are considered as raw material that describes a reality [1]. Any data that has restrictions and costs to access it is called closed data. Open data is data that anyone can freely use (reuse, distribute) for any purpose without the need for a license or permit. Open government data is a subset of open data that is considered as a key enabler for open government [2]. Danneels et al. have defined open government data as "accessible and reusable data services designed to allow third parties to create new value" [3]. Open government data is a concept that aims to involve citizens to the maximum extent in using or reusing government data. Citizen participation is expected to create innovation and help the government improve the quality of services provided [4]. In fact, open government data is an inseparable part of the concept of smart government [5].

The concept of open data is based on the idea that some data should be freely available to everyone so that they can use, reuse and publish it as they wish, without facing copyright, patent or other restrictions limitations [6]. The basic philosophy of open data is also similar to other movements based on the right of open access, such as open source software, open education, etc. Data, like any other commodity, has

potential benefits. Data must be refined like oil in order to reap its potential benefits; It means that it cannot be used by itself. When data becomes available to the public, individuals, organizations and scientists are able to recreate them in a new way and use them in the path of innovation and new value creation [7].

The purpose of this research is to understand, predict and control the situation of organizations that provide open government data. In order to understand the situation of organizations, comprehensive and effective criteria have been extracted in the quality of open government data based on library studies. Then, based on each criterion, the data of each organization has been extracted. Then, a classification model has been created in order to predict and control the status of the organizations.

2 RELATED WORKS

In this section, the research related to the evaluation of the quality of open government data has been discussed with the aim of understanding the current situation of organizations that provide open government data. Table 1 shows related researches based on researchers' names, year of publication, research title, brief description, and results.

Table 1: Research related to open government data quality assessment

Researchers' names	Year of publication	Research title	Brief description	Results	reference
Nikiforova et al	2021	Open government data portal usability: A user-centred usability analysis of 41 open government data portals	A set of 41 open government data portals has been selected for usability analysis according to the feedback of 40 users.	Based on the results of this research, the lack of interaction between users with open government data portals in cases such as providing feedback or requesting datasets is one of the main problems of open government data portals.	[8]
Zhang and Xiao	2020	Quality assessment framework for open government data: Meta-synthesis of qualitative research, 2009-2019	The integration of 10 qualitative studies in a common reference framework for evaluating the quality of government data has been addressed.	Based on a seven-step analysis, a common reference framework for evaluating the quality of open government data has been presented, which includes six criteria of accuracy, accessibility, completeness, timeliness, stability, and comprehensibility.	[9]
Nikiforova	2020	Timeliness of open data in open government data portals through pandemic-related data: a long data way from the publisher to the user	The study on the timeliness of open data in open government data portals has been studied.	60 countries and their portals are checked for timeliness of data.	[10]

de Juana-Espinosa and Luján-Mora	2020	Open government data portals in the European Union: A dataset from 2015 to 2017	They evaluated open government data portals in the European Union from 2015 to 2017. This study presents data collected from open government data portals in 28 EU countries.	In this research, the criteria of "the existence of a link from the open government data portal to the source site providing the dataset", "the existence of social network plugins" in order to discuss the experiences of users in using the open government data portal, "supporting different formats of the dataset" and "ability to search and filter data" have been suggested as evaluation criteria for open government data portals.	[11]
Kubler et al	2018	Comparison of metadata quality in open data portals using the Analytic Hierarchy Process	Providing quality criteria and checking their weight using multi-criteria decision-making techniques	It considers the following aspects in open data as quality dimensions: usability, completeness, openness, addressability and retrievability.	[12]
Dahbi et al	2018	Toward an evaluation model for open government data portals	They specify an evaluation model for open government data portals based on several main dimensions.	The specified dimensions are: information richness, discoverability, reusability and interactivity. The proposed evaluation model has been used to evaluate four national open government data portals.	[13]
Vetrò et al	2016	Open data quality measurement framework: Definition and application to Open Government Data	Present an approach for measuring the quality of open government datasets.	They suggest evaluating the data set for completeness, accuracy, traceability, comprehensibility, compliance, and expiration.	[14]
Dawes et al	2016	Planning and designing open government data programs: An ecosystem approach	Introduced a framework for evaluating the quality of open data portals at the national level and presented a set of criteria for evaluating data quality problems in open government data portals. These criteria were applied to 12 portals and several dimensions of data quality were introduced.	These dimensions included the existence of standards in data formats, the existence of metadata, the ability to read by machines and the up-to-datedness of the data.	[15]
Misuraca and Viscusi	2014	Digital governance in the public sector:	Have discussed a quality-based open government data	The criteria include three different quality dimensions:	[16]

		challenging the policy-maker's innovation dilemma	compliance assessment framework.	completeness, accuracy and timeliness.	
Ren and Glissmann	2012	Identifying information assets for open data: The role of business architecture and information quality	The quality criteria of government data have been identified.	Six open data quality criteria are proposed: accessibility and availability, comprehensibility, completeness, timeliness, error-free and secure.	[17]
Tauberer	2012	Open government data	It distinguishes optimal data quality from each other and defines 17 different characteristics that correspond to 5 categories of data quality.	The presented quality criteria are: basic principles, data format, universality of use, data dissemination and data openness.	[18]

As seen in previous researches, each of the conducted researches is focused on a specific aspect of improving the quality of open government data, and there is no comprehensive and complete set of evaluation criteria to understand the situation of organizations that provide open government data. Also, extracting the data of these organizations based on the extracted criteria and using these data in order to create a classification model to predict and control the status of organizations has not been done. In this research, after extracting the comprehensive quality criteria of open government data based on library studies, the data extraction of organizations that provide open government data has been done. Then, a classification model has been created to predict and control the state of organizations providing open government data.

3 MATERIALS AND METHODS

In this research, the type of research is based on the purpose of applied research. In the phase of understanding the situation of the organizations that provide open government data, effective criteria have been extracted for the quality of open government data and increasing citizens' willingness to use data based on library studies. Then the organizations, including all government organizations and institutions present in the open data portal that provide open government data, have been examined and the data related to each criterion has been calculated for each organization. The number of investigated government organizations and institutions was 112. Then, data mining techniques, including the classification model and the decision tree has been used to predict and control the status of organizations that provide open government data. Also, confusion matrix and criteria such as accuracy have been used to evaluate the created classification model. Figure 1 shows the diagram of the research process.

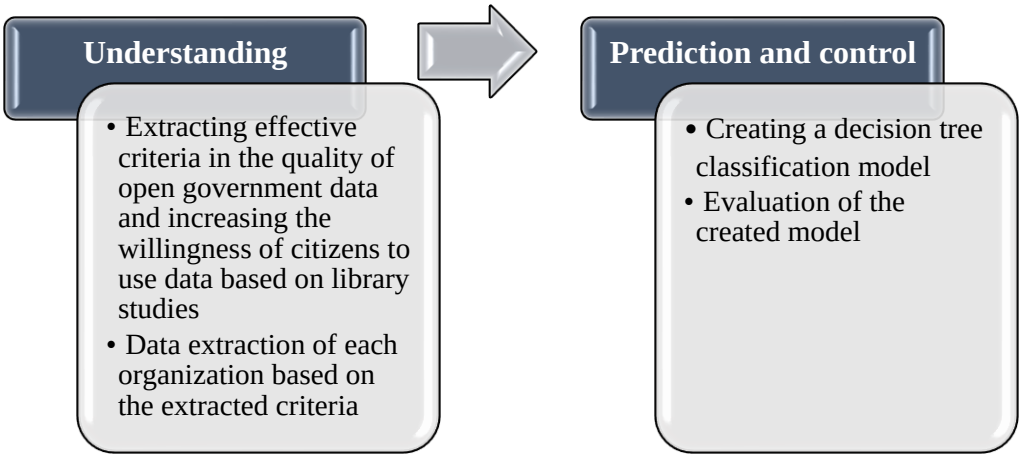

Figure 1: Diagram of the research process

4 RESULTS AND DISCUSSION

In this section, the results of the phases of understanding and predicting the situation of organizations that provide open government data are discussed.

4.1 Understanding the status of organizations that provide open government data

4.1.1 Extracting comprehensive and effective criteria for the quality of open government data and increasing the willingness of citizens to use the data

In order to understand the situation of organizations that provide open government data, comprehensive and effective criteria for the quality of open government data and increasing the willingness of citizens to use data have been extracted as evaluation criteria. For this purpose, library studies have been used. The extracted dimensions and criteria are stated in the following with the relevant reference.

4.1.1.1 Open data

1- Data accuracy [13, 19]: This dimension deals with data authenticity, absence of missing data, and up-to-datedness. In the following, the criteria related to the dimension of data accuracy are stated.

a) Data originality [14, 19]

b) Lack of missing data [13]

c) Up-to-datedness [13, 19, 20]: Each organization's score O_i is calculated based on the data it has published in the last five years (T_j is the data published in j years ago) using formula 1.

$$O_i = \sum_{j=0}^{4}(1 - \frac{j}{10})\times T_j \quad (1)$$

2- Discoverability [13]: Discoverability deals with tools and mechanisms that increase data accessibility and browsing. In the following, the criteria related to the dimension of discoverability are stated.

a) Metadata completeness [13, 20-22]: The checked fields are: title, description, label, publisher, etc.

b) Data access [13, 14, 20, 22]: This criterion assesses the presence of features that enhances data discovery, specifically the presence of three features: search, sort, and filter.

3- Richness of information [13]: Information richness examines the satisfaction of user needs in terms of the amount of data. In the following, the criteria related to the dimension of information richness are stated.

a) Number of data sets [13, 14, 20, 21]
b) Number of categories of data sets [21]
c) Data subject matter [21]
d) Data request ability [13, 22]

4.1.1.2 Data transparency

1- Reusability [13, 19]: Open government data is considered reusable when the data is released under an open license that permits unrestricted access, reuse, and redistribution of the data. It should also be published in electronic format and be machine-readable. In the following, the criteria related to the dimension of reusability are stated.

a) License openness [13, 19]: This criterion evaluates the openness of the dataset license for reuse.
b) Format openness [13, 19, 21]: For each dataset D_n, the FOI_n score is assigned based on the source format as follows:
    - If the format is not machine-readable: $FOI_n = 0$ (e.g. PDF)
    - If the format is machine-readable: $FOI_n = 1$ (e.g. JSON or CSV)
c) Free [19]
d) Non-discriminatory [14, 19]: the possibility of accessing and reusing data should be the same for all people.

2- Understandable [19]

4.1.1.3 Interactivity

1- Feedback [13, 19, 20]: This criterion measures the presence of three possibilities: commenting on the dataset, ranking the dataset, feedback on the portal.

2- Visualization [13, 20, 21]: This measure assesses the presence of visualization tools and features such as maps, charts, or programs to visualize and interact with data.

4.1.2 Data extraction of organizations based on extracted criteria

After extracting the criteria in the previous phase, the data of the organizations and government institutions present in the open data portal was extracted based on the extracted criteria. The number of organizations in this portal that provide open data has been 112.

4.2 Predicting and controlling the status of organizations that provide open government data

In this phase, data mining techniques including classification have been used in order to predict and control the status of organizations that provide open government data. These techniques have been applied on the data obtained in the previous phase.

4.2.1 Classification model

In this phase, the decision tree classification model was created using Rapid miner software. Figure 2 shows a view of the implementation of this classifier and the operators used.

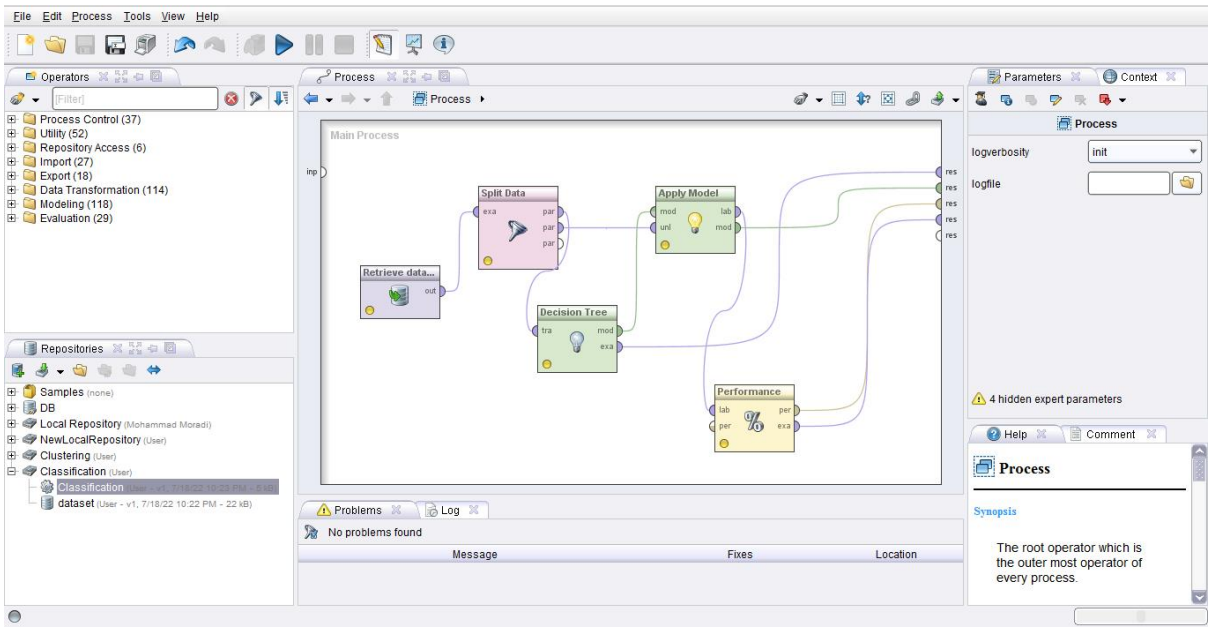

Figure 2: A view of the implementation of the decision tree classification model in the Rapid miner software

In this classification, 80% of the data were used as training data and the remaining 20% were considered as test data. Figure 3 shows the created decision tree. "HV" is high visitation, "MV" is medium visitation and "LV" is low visitation.

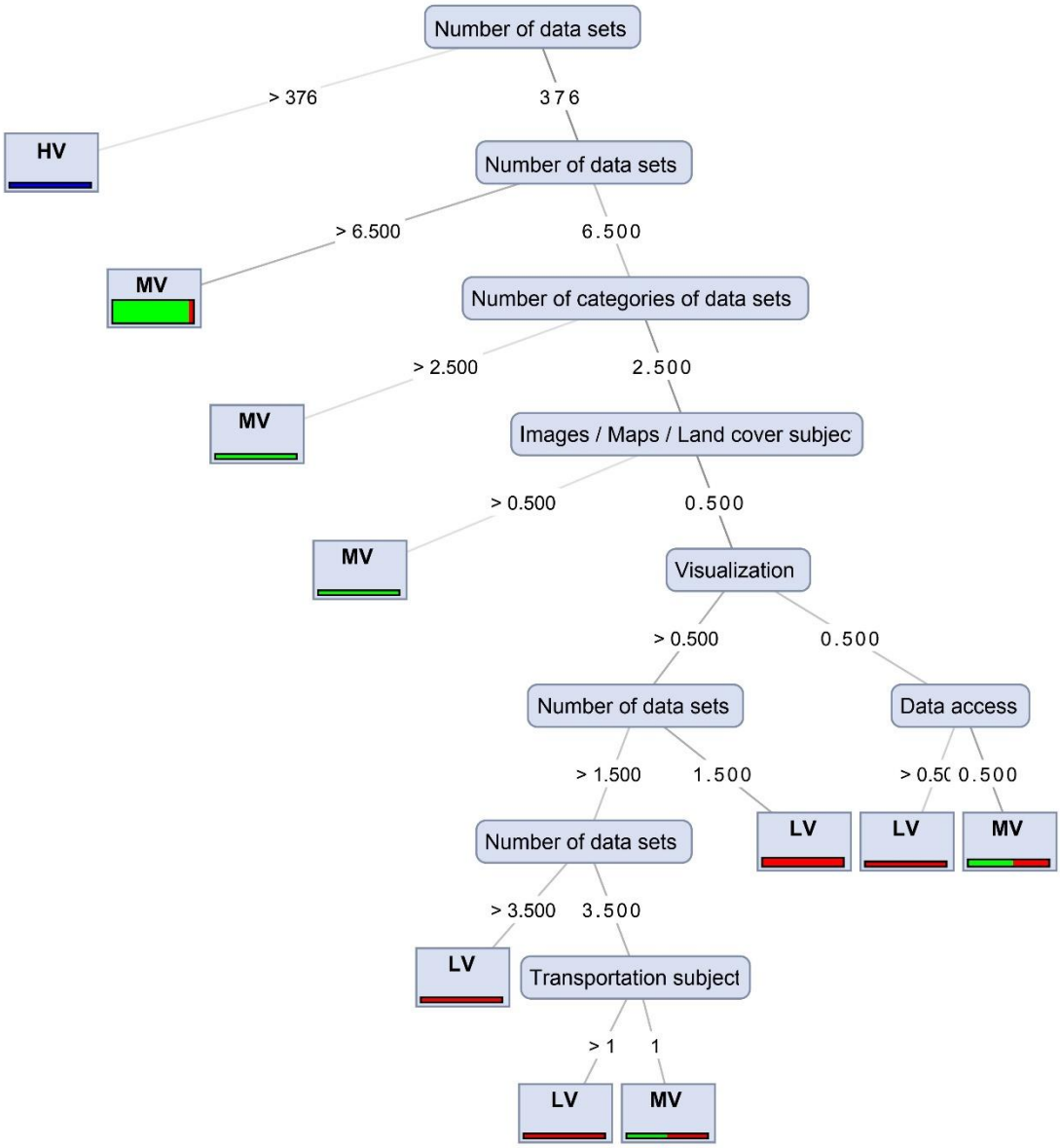

Figure 3: Decision tree

Using the decision tree created in Figure 3, it is possible to predict the status of organizations in providing open government data. For example, if the number of data sets is less than 7 and the number of categories of datasets is greater than 2, the organization status will be equal to MV i.e. medium visitation. Also, by using the created decision tree, it is possible to control the status of organizations and improve their status. For example, if the status of an organization is equal to the low visitation (LV), it can be changed to the medium visit status by increasing the number of data sets, increasing visualization and data access. The amount of changes required for each mentioned criterion can also be obtained from this decision tree.

4.2.2 Evaluation of the classification model

In this section, the evaluation of the created classification model has been discussed. Figure 4 shows the confusion matrix of the created decision tree.

accuracy: 90.91%

	true HV	true MV	true LV	class precision
pred. HV	0	0	0	0.00%
pred. MV	0	14	1	93.33%
pred. LV	0	1	6	85.71%
class recall	0.00%	93.33%	85.71%	

Figure 4: Confusion matrix of the created decision tree

The accuracy of the created classification model is equal to 90.91%, which shows that the model is desirable. Also, the values of other criteria such as recall and precision are shown in Figure 4.

5 CONCLUSION

In order to increase citizens' desire for open government data and take advantage of its benefits, it is necessary to first understand the complex system of communication between citizens and organizations that provide open government data and determine the status of each organization. For this purpose, a series of evaluation criteria is needed. In this research, comprehensive and complete criteria for evaluating the status of organizations were presented. Also, the data of the organizations present in the open data portal were extracted based on each criterion. Then, a decision tree classification model was presented in order to predict and control the status of organizations. The accuracy of the created model was equal to 90.91, which shows that the model is desirable. By using the presented classification model, managers will be able to predict the status of organizations. Also, they can apply the necessary changes on the organizations in order to change the status of the organization to the desired status.

REFERENCES

[1] Pejić Bach, Mirjana, Tine Bertoncel, Maja Meško, Dalia Suša Vugec, and Lucija Ivančić. "Big data usage in european countries: Cluster analysis approach." *Data* 5, no. 1 (2020): 25.

[2] Attard, Judie, Fabrizio Orlandi, Simon Scerri, and Sören Auer. "A systematic review of open government data initiatives." *Government information quarterly* 32, no. 4 (2015): 399-418.

[3] Danneels, Lieselot, Stijn Viaene, and Joachim Van den Bergh. "Open data platforms: Discussing alternative knowledge epistemologies." *Government Information Quarterly* 34, no. 3 (2017): 365-378.

[4] Zuiderwijk, Anneke, Rhythima Shinde, and Marijn Janssen. "Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results?." *International review of administrative sciences* 85, no. 4 (2019): 645-672.

[5] Arief, Assaf, and Dana Indra Sensuse. "Designing A Conceptual Model for Smart Government in Indonesia using Delphi 2 nd Round Validity." In *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 93-98. IEEE, (2018).

[6] Wiegand, Viola. "Book review: Markus Rheindorf, Revisiting the Toolbox of Discourse Studies: New Trajectories in Methodology, Open Data, and Visualization." (2021): 102-104.

[7] Enders, Tobias, Carina Benz, and Gerhard Satzger. "Untangling the open data value paradox: How organizations benefit from revealing data." In *International Conference on Wirtschaftsinformatik*, pp. 200-205. Springer, Cham, (2021).

[8] Nikiforova, Anastasija, and Keegan McBride. "Open government data portal usability: A user-centred usability analysis of 41 open government data portals." *Telematics and Informatics* 58 (2021): 101539.

[9] Zhang, Hui, and Jianying Xiao. "Quality assessment framework for open government data: Meta-synthesis of qualitative research, 2009-2019." *The Electronic Library* 38, no. 2 (2020): 209-222.

[10] Nikiforova, Anastasija. "Timeliness of open data in open government data portals through pandemic-related data: a long data way from the publisher to the user." In *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*, pp. 131-138. IEEE, (2020).

[11] de Juana-Espinosa, Susana, and Sergio Luján-Mora. "Open government data portals in the European Union: A dataset from 2015 to 2017." *Data in brief* 29 (2020): 105156.

[12] Kubler, Sylvain, Jerermy Robert, Sebastian Neumaier, Jürgen Umbrich, and Yves Le Traon. "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process." *Government Information Quarterly* 35, no. 1 (2018): 13-29.

[13] Dahbi, Kawtar Younsi, Hind Lamharhar, and Dalila Chiadmi. "Toward an evaluation model for open government data portals." In *International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*, pp. 502-511. Springer, Cham, (2018).

[14] Vetrò, Antonio, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, and Federico Morando. "Open data quality measurement framework: Definition and application to Open Government Data." *Government Information Quarterly* 33, no. 2 (2016): 325-337.

[15] Dawes, Sharon S., Lyudmila Vidiasova, and Olga Parkhimovich. "Planning and designing open government data programs: An ecosystem approach." *Government Information Quarterly* 33, no. 1 (2016): 15-27.

[16] Misuraca, Gianluca, and Gianluigi Viscusi. "Digital governance in the public sector: challenging the policy-maker's innovation dilemma." In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, pp. 146-154. (2014).

[17] Ren, Guang-Jie, and Susanne Glissmann. "Identifying information assets for open data: The role of business architecture and information quality." In *2012 IEEE 14th International Conference on Commerce and Enterprise Computing*, pp. 94-100. IEEE, (2012).

[18] Tauberer, Joshua. *Open government data*. Joshua Tauberer, (2012).

[19] Veljković, Nataša, Sanja Bogdanović-Dinić, and Leonid Stoimenov. "Benchmarking open government: An open data perspective." *Government Information Quarterly* 31, no. 2 (2014): 278-290.

[20] Huang, Ruhua, Chunying Wang, Xiaoyu Zhang, Dan Wu, and Qingwen Xie. "Design, develop and evaluate an open government data platform: a user-centred approach." *The Electronic Library* (2019).

[21] Saxena, Stuti. "Open government data (OGD) in six Middle East countries: An evaluation of the national open data portals." *Digital Policy, Regulation and Governance* (2018).

[22] Zheng, Lei, Wai-Min Kwok, Vincenzo Aquaro, Xinyu Qi, and Wenzeng Lyu. "Evaluating global open government data: Methods and status." In *Proceedings of the 13th international conference on theory and practice of electronic governance*, pp. 381-391. (2020).