Analysing User Involvement in Open Government Data Initiatives

Dagoberto Jose Herrera-Murillo[0000−0002−8000−628X], Abdul Aziz[0000−0003−3615−4573], Javier Nogueras-Iso[0000−0002−1279−0367], and Francisco J. Lopez-Pellicer[0000−0001−6491−7430]

Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Spain
{dherrera,abdul.aziz,jnog,fjlopez}@unizar.es
https://www.iaaa.es

Abstract. Over the last decade, many Open Data initiatives have been launched by public administrations to promote transparency and reuse of data. However, it is not easy to assess the impact of data availability from the perspective of user communities. Although some Open Data portals provide mechanisms for user feedback through dedicated discussion forums, web forms, and some of the user experiences are listed as use cases in their portals, there is no consistent way to compare user feedback in different data initiatives. To overcome the difficulty of assessing user impact, this paper examines the activity generated by Open Data initiatives through the social network Twitter: a forum used by all types of stakeholders and publicly available for consistent analysis. We propose a methodology to compile a set of variables that describe both the main characteristics of Open Data initiatives and the associated Twitter activity. The collected data is then analysed using factor analysis and clustering techniques to derive possible relationships between the variables. Finally, the initiatives are classified according to their activity on social networks and the values that characterise some of their features. The methodology was evaluated by analysing 27 European Open Government Data portals and their activity on Twitter in 2021.

Keywords: Open Government Data · Open Data Portals · Metadata Quality · User Engagement · Social Media

1 Introduction

In the current digital world, the movement of Open Data (OD) is expanding at a breakneck speed, and the ever-increasing availability of data at Open Data Portals is fuelling the expansion of this movement [19, 6]. Governments are increasingly implementing open data projects and setting up open data portals to facilitate the distribution of this data in open and reusable formats. As a result, a vast number of open data repositories, catalogues, and websites have sprouted up. The philosophy of openness in Open Data is to use, share and access the data freely in any format. Open Data portals are online catalogues that contain

dataset descriptions, i.e. they are a type of digital library. Such catalogues allow the discovery and management of metadata records describing datasets which maybe available for access or download in one or more distribution formats. Governments acquire and generate massive volumes of data. In addition, metadata records describe datasets in terms of authorship, provenance, and licensing, among many other aspects [13].

According to the European Commission [5], Open Government Data (OGD) portals play a vital role in opening the data and the continuous publication of open data in OGD portals raises the demand for high-quality data and the quality of the portal itself. In this respect, the use and reuse of public sector data is a significant factor in driving the current trend of opening government data through the EU portal [5, 20].

Most of the current OD ecosystems are not user-driven and do not adequately match supply and demand. It is generally accepted that the role of users is critical for the development of OD ecosystems, but the current ecosystems are driven by providers [22]. Engagement is a critical success factor to make current open OGD initiatives more user-centric. The lack of mature feedback and interaction mechanisms to engage users, however, can be seen as a major limitation. Focusing directly on receiving feedback from users, the portals of some Open Data initiatives have proposed communication channels with users through dedicated discussion forums or web forms where different communities of users can report on their experiences of reusing data made available through the portals. Some of these initiatives even provide specialised tools to access the data and offer storytelling features to users [1]. However, this kind of user feedback is very heterogeneous and feedback from different initiatives cannot be compared automatically.

This is the main reason for our decision to study user feedback in social networks: social networks are a general forum where different stakeholders express their opinions about any kind of activity or organisation. Moreover, social media can play a strategic role in improving visibility by encouraging users to visit the portal and engaging them by presenting the available data and portal features [18].

Twitter is the most attractive social platform when it comes to measuring user engagement of open data portals. Among the many reasons why this platform is useful are the following: Twitter is one of the social media platforms with the largest audience; it is used to discuss topics ranging from personal to professional interests; and, at least in Europe, it is the most widely used social media channel by open government data initiatives [18]. In contrast to other digital social networks, it has a greater tendency to circulate academic content and knowledge [9].

The purpose of this work is to analyse user involvement in Open Government Data initiatives. To this end, we have compiled a set of variables that describe both the main characteristics of the Open Data initiatives and the associated Twitter activities. These are later analysed through factor analysis and clustering techniques to derive possible relationships between the variables and the

classification of the initiatives according to their activity on the social networks, as well as the values that characterise some of their features.

The rest of the paper is as follows. Section 2 provides a review of the relevant literature, where we also discuss the methodologies used in the past for Open Data initiatives. The methodology that includes the analytic framework of the working model is presented in Section 3. Section 4 presents the findings and results of this research. Section 5 compares our findings with the Open Data Maturity Report for 2021. We conclude with a summary of the contributions and future work.

2 Related Work

There are several research works in the literature about the monitoring of the quality of Open Data Portals [11, 15], which are relevant to have an overall perspective of the current status of Open Data initiatives, their maturity or their commitment to FAIR principles [23], but they did not take into account any insights of the direct opinion of user engagement [24]. Likewise, Begany and Gil-Garcia [3] monitored the levels of user engagement by analysing web analytic behavioural data taken from the New York State open health data portal. In addition, they emphasised the actual use of open data and more specifically how users of Open Data Portals interact with open datasets.

Concerning the study of influence in social networks, several research works in the literature have investigated how to measure the impact of organisations and Twitter profiles. For instance, Berrocal et al. [4] studied the influence of University Libraries on Twitter using an influence index based on Klout [7]. Furthermore, Khan et al. [10] explored data citation and reuse practices in 43,802 openly available biodiversity datasets. The altmetrics sourced from blogs, Twitter, Facebook, and Wikipedia suggest that social activity is driven by data publishers and data creators. Authors made a hypothesis that such activities are promotion-related and may lead to more reuse of open datasets.

Concerning user-centric Open Government Data Initiatives, Nikiforova and McBrid [14] analysed and compared the various contexts regarding the employment of Open Government Data Portals by users and emphasising the most often disregarded user-centred aspects. They used the questionnaire technique to verify the user-centric usability of Open Government Data Portals. Notably, Zhu and Freeman[24] evaluated different methods of user interactions with Open Government Data Initiatives and developed a framework called user interaction framework where they evaluated the U.S. Municipal Open Data Portals and provided the findings regarding user understanding and engagement with the data portals.

For several years the Open Data Maturity Report [18] has benchmarked the development of European countries in the field of open data. The document mentions four dimensions: policy, impact, portal, and quality. In the portal dimension, it includes a sustainability variable that identifies actions conducted to ensure the portal visibility, including social media presence. According to

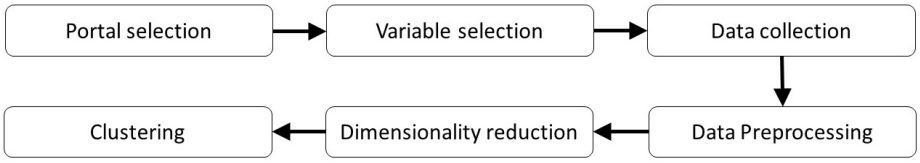

Fig. 1: Proposed Methodology for Data Processing

the Open Data maturity report, Twitter is the most widely used social media channel in 16 of the analysed countries.

3 Methodology

This research utilises a quantitative approach for analysing multiple metrics related to the national Open Data Portals of EU member countries. As shown in Figure 1, the proposed methodology consists of 6 steps, which are described below:

Portal Selection: The first step of the methodology is to find appropriate resources to identify the location of portals and the documents describing their features. The list of platforms studied comes from two sources: the national catalogues of the European Data Portal (EDP) [16], and the compilation made by Juana-Espinosa and Lujan-Mora [6]. Both sources showed a high degree of concordance.

Variable Selection: This refers to the process of choosing relevant variables describing the features of the Open Data portal to include in our model. The relevant variables and their sources are shown in Table 1. In our experimental design, we make a choice of variables taking advantage of the sources of information available through the national portals under observation and through the European Data Portal. Therefore, some of the most representative operational attributes are gathered from the EDP (ND, ODM, MQA, URL), some of them are Twitter activity metrics (NT, TFP, UT, NI), and there are other variables (NU, GS) that help us understand the magnitude of data reuse around each portal. It must be noted that the variables present in Table 1 are the final selection of variables: our experiment included some other variables and combinations of them, but they were discarded due to their negative effect on the feasibility of indicators obtained during experiment in the last two phases of the methodology (Dimensionality Reduction and Clustering). Variables present in Table 1 are by no means intended as a complete and exhaustive list. In fact, later steps help us to explore the underlying structure that may be useful for refining the variable selection in the future.

Data Collection: This refers to the process of gathering data from reliable sources mentioned in Table 1 that guarantee the reproducibility of the measurements. We assume that the values obtained are valid and representative as they are gathered from recognized sources such as the European Data Portal and

Table 1: Description of the variables

Variable	Description	Source
ND	Number of datasets available for consultation	Automatic from data.europa.eu
ODM	Open Data Maturity score (0-100)	Manual from data.europa.eu
MQA	Metadata Quality Assurance rating (0-405)	Automatic from data.europa.eu
URL	% of accessible URLs	Automatic from data.europa.eu
NU	Number of data use cases listed in the portal	Manual from portals
GS	Number of items in Google Scholar citing the portal	Manual from Google Scholar
NT	Number of relevant Tweets	Automatic, derived from Twitter API
TFP	Number of Tweets by portal account	Automatic Twitter API
UT	Number of users posting Tweets	Automatic Twitter API
NI	Number of interactions generated by Tweets. This corresponds to the sum of retweets, replies, quotes and likes.	Automatic Twitter API

the academic Twitter API. The variables representative of the EDP can be collected through the EDP API (MQA, ND, URL) [17] or manually (ODM). The variables measuring the conversation on Twitter related to portals for the year 2021 (NT, TFP, UT, NI) can be collected using the Twitter API for Academic Research [21]. This API allows the retrieval of tweets whose text mentions the URL of portals or their Twitter accounts. Finally, the number of use cases listed in a data portal (NU) and the number of mentions in Google Scholar (GS) must be collected manually for each data portal.

Data Processing: This consists of preparing the raw data and making it suitable for the analytical models. First, we must compute the correlations between the metrics using the Spearman coefficient. This coefficient can range from -1 to 1, with -1 or 1 indicating a perfect monotonic relationship: when the value of one variable increases, the other variable value also increases or decreases. After that, we must normalise the variables by removing the mean and scaling them to unit variance.

Dimensionality Reduction: This step involves exploring the underlying variable structure and reducing the data to a smaller number of explainable factors. For this purpose, we propose the use of factor analysis to reduce the dimensions of the original dataset [8]. Likewise, Bartlett test ($X^2 = 166.56$, $p < 0.0001$) and the Kaiser–Mayer–Olkin test ($KMO = 0.67$) are employed to verify the feasibility of the overall factor analysis. We take into account the Kaiser-Guttman criterion ($eigenvalue > 1.0$) to decide the optimal number of factors and, for each factor, only variables with loading greater than 0.4 after applying Varimax rotation are considered to influence the factor.

Clustering: Clustering consists of grouping portals into groups based on the dimensions that describe them. For this step, we propose to apply three common clustering methods: hierarchical clustering, K-means clustering, and K-medians clustering [12] (less sensitive to outliers). Combining these clustering techniques is a common way to improve the robustness of the final results [2]. The ideal number of clusters for K-means is defined by plotting the explained variation as a function of the number of clusters and identifying the inflection point at which adding another cluster does not improve much better intra-cluster variation, a procedure also known as the "elbow method".

4 Results

This section displays the outcomes of applying the proposed methodology on the national Open Data Portals of 27 EU member countries and their Twitter activity in 2021. We selected 2021 as this is the last year with complete information on Twitter activity. In addition, the values obtained from the Open Data Maturity report or from available APIs also reflect the situation after year 2021 had finished (when the experiment was performed).

Table 2 shows the results about the values of variables for the 27 portals under observation with the mean and coefficient of variation (CV) corresponding to each of them. The variables describing Twitter activity (NT, TFP, UT, NI) and the number of use cases (NU) are the ones with the greatest relative variability. Similarly, in terms of relevance to portals, France, Spain, and Austria have the highest nominal values for the parameters of Twitter conversation, number of use cases, and Google Scholar mentions. The Hungarian platform is the only one that does not follow a catalogue structure and does not have values for most of the indicators under observation.

Furthermore, the Spearman rank correlation coefficient is used to measure the strength and direction of association between pairs of variables, which is shown in Table 3. While looking at the Twitter metrics (except for the number of tweets by the portal account itself), the number of use cases and mentions in Google Scholar are strongly and positively correlated with each other. Moreover, the Metadata Quality Assurance rating correlates positively and moderately with the number of datasets (0.54) and the percentage of accessible URLs (0.54). The remaining correlations are weak.

Given the high correlation between the Twitter conversation variables, we removed UT and NI before factor analysis to reduce the effect of multicollinearity. The outcome for a three-factor solution accounting for 72% of the variance is shown in Table 4. The number of use cases, mentions in Google Scholar, and tweets are the variables that best explain factor 1. The number of datasets, MQA rating, and the percentage of accessible URLs are the most representative variables for factor 2. Finally, the number of tweets by portal account is considered the best variable for factor 3, with a small contribution of the number of tweets. The ODM score did not load in any of the three factors. From the factor loadings, factor scores are computed for each portal.

Table 2: Values of variables for Open Government Data portals of the EU countries and their Twitter activity in 2021

Country*	Portal URL	ND	ODM	MQA	URL	NU	GS	NT	TFP	UT	NI
FR	data.gouv.fr	41,881	98	172	67	3,099	556	1,843	45	921	33,750
ES	datos.gob.es	60,102	95	196	46	400	137	1,384	448	294	9,974
AT	data.gv.at	38,586	92	199	93	689	158	258	110	85	3,696
IT	dati.gov.it	53,490	92	152	54	0	31	214	44	35	1,041
IE	data.gov.ie	13,815	95	185	42	23	73	173	3	46	1,377
LV	data.gov.lv	612	77	165	49	0	19	144	0	30	1,914
PL	dane.gov.pl	26,180	95	166	99	45	104	116	0	57	1,481
LU	data.public.lu	1,613	66	131	97	150	24	104	37	25	319
NL	data.overheid.nl	21,259	92	192	89	118	53	95	40	40	363
DE	govdata.de	51,275	89	240	56	24	118	85	9	55	1,502
CZ	data.gov.cz	142,554	74	276	99	0	19	62	43	11	702
GR	data.gov.gr	10,446	82	106	29	0	36	44	0	37	303
BG	data.egov.bg	10,680	78	47	0	0	6	37	15	12	119
FI	avoindata.fi	2,058	86	203	4	77	60	28	3	17	571
RO	data.gov.ro	2,753	76	98	7	10	15	28	0	18	24
DK	opendata.dk	823	91	164	42	0	22	27	14	12	137
PT	dados.gov.pt	4,928	66	183	81	51	43	25	0	14	242
CY	data.gov.cy	1,210	91	226	12	47	9	8	3	6	52
SE	www.dataportal.se	7,825	84	170	30	0	15	8	0	6	153
HR	data.gov.hr	1,141	84	96	52	6	22	6	0	3	23
BE	data.gov.be	13,056	55	218	31	81	12	3	0	3	15
SI	podatki.gov.si	5,098	92	120	61	14	17	2	0	2	12
EE	avaandmed.eesti.ee	879	94	0	0	150	5	0	0	0	0
LT	data.gov.lt	1,721	89	99	56	28	3	0	0	0	0
MT	open.data.gov.mt	205	51	0	0	0	2	0	0	0	0
SK	data.gov.sk	2,862	50	124	0	11	12	0	0	0	0
HU	kozadat.hu	0	58	0	0	0	0	0	0	0	0
Mean		19150.1	81.2	145.5	44.3	186.0	58.2	173.9	30.1	64.0	2139.6
CV		1.6	0.2	0.5	0.8	3.2	1.9	2.4	2.9	2.8	3.1

*We are using the two letter-code of ISO-639 to refer to the country of the Open Data initiatives that have been analysed.

Table 3: Spearman correlation

	ND	ODM	MQA	URL	NU	GS	NT	TFP	UT	NI
ND	1									
ODM	0.19	1								
MQA	0.54**	0.30	1							
URL	0.49*	0.32	0.54**	1						
NU	0.19	0.29	0.12	0.21	1					
GS	0.28	0.39*	0.26	0.30	0.96**	1				
NT	0.33	0.36	0.20	0.20	0.84**	0.87**	1			
TFP	0.39*	0.26	0.23	0.17	0.19	0.26	0.63**	1		
UT	0.25	0.33	0.16	0.19	0.97**	0.96**	0.94**	0.33	1	
NI	0.26	0.32	0.16	0.19	0.97**	0.96**	0.93**	0.32	1.00**	1

** $p < 0.01$ and *$p < 0.05$ indicate significant correlation.

Table 4: Rotated matrix for factor analysis

Variable	Factor		
	1	2	3
ND		0.64	
ODM			
MQA		0.78	
URL		0.72	
NU	0.97		
GS	0.96		
NT	0.84		0.52
TFP			0.94
Eigenvalues	2.70	1.80	1.30
Variance	0.34	0.22	0.16
Cum. Variance	0.34	0.56	0.72

Note: Loadings with absolute values below 0.40 are omitted from the table

Observing high factor loadings associated with particular variables implies that these variables contribute more to this component. Therefore, portals with high values on these variables tend to have higher factor scores on this particular dimension and vice versa for low values.

The next stage of the research process involved clustering methods to group the EU data portals. Figure 2 shows the clustering dendrogram, which is the result of the hierarchical clustering algorithm. In addition, Figure 3 shows the best clustering solution for k-means (k = 5) using a cluster profiling plot in parallel coordinates (the clustering profiling plot obtained with K-medians is almost identical). Parallel coordinates are a frequent way of visualising how the Open Government Data Initiatives differ from each other across factors.

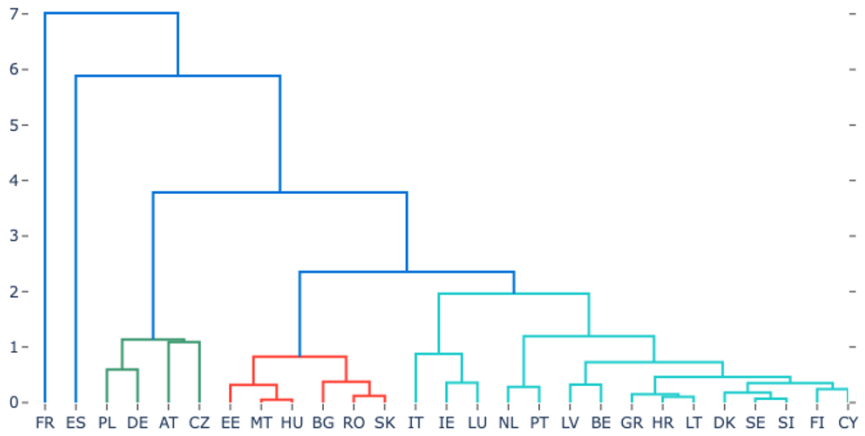

Fig. 2: Cluster dendrogram.

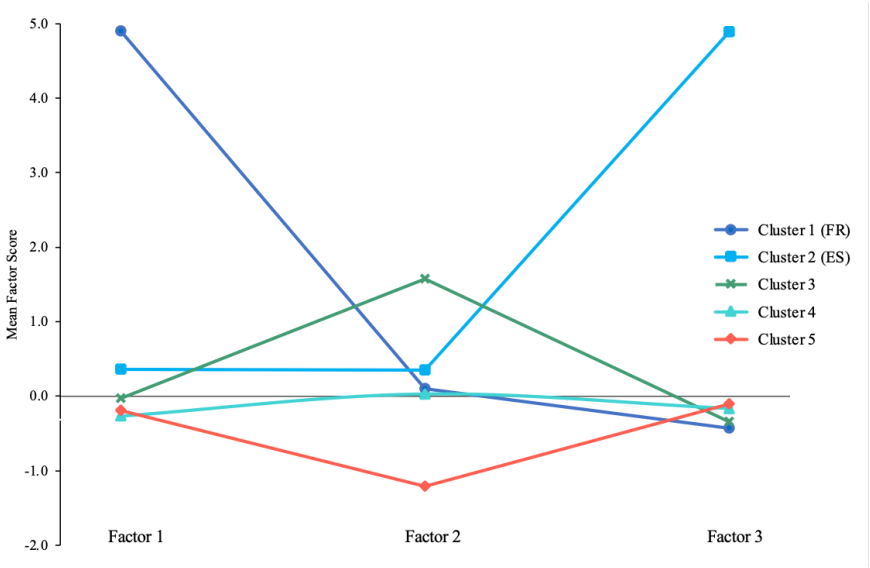

Fig. 3: Cluster profile plot for mean factor score of the clusters obtained with K-means.

Moreover, the composition of Open Data initiatives that form each cluster for K-means, K-medians and hierarchical clustering is shown in Table 5. In general, the techniques converge for the identification of 5 groups that are clearly distinguished from each other based on their behaviour through the factors. The resulting assignments of hierarchical clustering (with a cutoff distance equal to 2) and K-medians were identical. K-means reallocated one member (NL).

The first cluster is determined by the preeminent performance of France in factor 1. In the second cluster, Spain stands out for its score in factor 3, and has a slightly higher value for factor 1. Cluster 4 is the most numerous and shows an intermediate behaviour in the three factors. Cluster 3 and cluster 5 are characterised respectively by high and low values of the variables in factor 2.

Table 5: Cluster membership

Cluster	K-means	K-medians	Hierarchical Clustering
1	FR	FR	FR
2	ES	ES	ES
3	AT, PL, NL, DE, CZ	AT, PL, DE, CZ	AT, PL, DE, CZ
4	IT, IE, LV, LU, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT	IT, IE, LV, LU, NL, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT	IT, IE, LV, LU, NL, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT
5	BG, RO, EE, MT, SK, HU	BG, RO, EE, MT, SK, HU	BG, RO, EE, MT, SK, HU

5 Discussion

Our methodology was assessed with the analysis of 27 European Open Government Data initiatives and the Twitter activity generated in 2021. This allows us to compare our findings with the Open Data Maturity Report for the year 2021.

For our experiment, we used correlation analysis and factor analysis, which indicate the existence of a dimensional structure. On the one hand, the activity on Twitter, the number of use cases, and the mentions in Google Scholar point to a dimension that we could call "user community drivenness". On the other hand, the indicators of metadata quality and quantity of datasets describe a dimension that we could call "data compliance drivenness". In addition, the number of tweets from the portal suggests that the promotion given by the portals themselves to the distribution of content could characterise a dimension of "portal community drivenness".

The cluster analysis allows us to profile the European national portals based on the previously identified dimensions. First, we observe how the French Open Data initiative can be defined as "user community driven". This result is consistent with what the Open Data Maturity Report 2021 indicates, where France is the best-positioned country and is described as user-centric. The main reason for this can be the efforts paid by the French Open Data initiative that monitors the user feedback through multiple functionalities, including discussion forums on the individual datasets. Second, we see that the Spanish Open Government Data initiative has a high Twitter activity but this activity is mainly ruled by the public body coordinating the initiative. The Open Data Maturity Report highlights the efforts of this portal to create editorial content, optimise the search and discoverability of content, and use actively Facebook, LinkedIn, YouTube, and Flickr. Spain also reports using social listening tools, web analytics, and SEO positioning. Third, there are some initiatives with remarkable quality and size of published datasets, which do not hold a direct impact on Twitter activity, probably because the bodies coordinating the initiatives are not so active in disseminating their work in social networks. In general, these countries are positioned in the best performing categories of the Open Data Maturity Report. Fourth, we can observe that most initiatives report a medium-low level of quality and social network activity. Last, there are also some initiatives with the lowest level of quality and almost no presence on Twitter, which probably denotes that they have been started recently and are not mature enough to generate the interest of users. The members of this cluster correspond to the lagging categories of the Open Data Maturity Report.

Given the exploratory nature of the study, we would like to reflect on a series of issues that could serve to improve and deepen the measurement of user involvement in Open Government Data Initiatives. One of these issues is the effect of using absolute values of variables instead of relative values. In this regard, the population of the countries or the number of published datasets could be used as weighting factors of the involved variables. However, in the experiment carried out, these variables were poorly correlated with all the others. For instance, while populous countries like France and Spain nominally lead the interactions,

we can see at the same time cases like Germany performing modestly for its large population. At this point, we can only formulate hypotheses to explain this lack of direct correlation between the population of the reference country and the volume of interactions around its national open data portal. Open data could not be considered a mainstream phenomenon and be limited to very compact community niches where the size of the national reference population has a secondary effect.

6 Conclusion

To conclude, this paper has proposed a methodology for measuring user involvement in Open Data initiatives by analysing the activity generated on Twitter and trying to understand the relationship between the social network activity and the main features characterising the size, quality, and maturity of Open Data initiatives. Moreover, apart from compiling the values of the different selected variables for these initiatives, there are relevant conclusions that can be derived from the results obtained through factor analysis and clustering techniques. Overall, policy makers can use findings to benchmark Open Data initiatives according to their interaction with the user community.

As future work, we would like to perform additional experiments validating the methodology and include a temporal analysis of the evolution of Twitter activity generated by Open Data initiatives since the year of their launch. Moreover, we would also like to explore the potential of making a qualitative analysis of the content through the use of techniques for sentiment analysis and semantic analysis of the tweets mentioning the Open Government Data Initiatives.

Last, in line with the growing interest in monitoring and measuring the open data re-use and the impact it generates, we hope that our work stimulates the discussion on the development of quantitative and qualitative alternative metrics for the evaluation of the impact of Open Data initiatives.

Acknowledgements

This paper is partially supported by the Aragon Regional Government through the project T59_20R. The work of Dagoberto José Herrera-Murillo and Abdul Aziz is supported by the ODECO project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 955569.

References

1. Lorea Akerreta Escribano and Julián Moyano Collado. Contar historias con los datos: Aragón Open Data Focus, una experiencia innovadora de reutilización de los datos del sector público. *Scire: representación y organización del conocimiento*, 27(1):31–43, 2021.

2. Margarita Argüelles, María del Carmen Benavides, and I Fernández. A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Applied Economics*, 46(21):2511–2519, 2014.
3. Grace M Begany and J Ramon Gil-Garcia. Understanding the actual use of open data: Levels of engagement and how they are related. *Telematics and Informatics*, 63:101673, 2021.
4. José Luis Alonso Berrocal, Carlos G Figuerola, and Ángel F Zazo Rodriguez. Propuesta de índice de influencia de contenidos (Influ@ RT) en Twitter. *Scire: representación y organización del conocimiento*, pages 21–26, 2015.
5. Wendy Carrara, Wae-San Chan, Sander Fischer, and Eva Van-Steenbergen. Creating value through open data: Study on the impact of re-use of public data resources. European Commission, 2015.
6. Susana de Juana-Espinosa and Sergio Luján-Mora. Open government data portals in the European Union: A dataset from 2015 to 2017. *Data in brief*, 29:105156, 2020.
7. Chad Edwards, Patric R Spence, Christina J Gentile, America Edwards, and Autumn Edwards. How much Klout do you have... A test of system generated cues on source credibility. *Computers in Human Behavior*, 29(5):A12–A16, 2013.
8. Joseph Hair, Barry Balbin, William Black, and Rolph Anderson. *Multivariate Data Analysis*. Cengage Learning EMEA, 2019.
9. Stefanie Haustein, Rodrigo Costas, and Vincent Larivière. Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PloS one*, 10(3):e0120495, 2015.
10. Nushrat Khan, Mike Thelwall, and Kayvan Kousha. Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics. *Scientometrics*, 126(4):3621–3639, 2021.
11. Sylvain Kubler, 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(1):13–29, 2018.
12. Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In *International conference on machine learning*, pages 7055–7065. PMLR, 2020.
13. Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality (JDIQ)*, 8(1):1–29, 2016.
14. Anastasija Nikiforova and Keegan McBride. Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, 58:101539, 2021.
15. Javier Nogueras-Iso, Javier Lacasta, Manuel Antonio Ureña-Cámara, and Francisco Javier Ariza-López. Quality of Metadata in Open Data Portals. *IEEE Access*, 9:60364–60382, 2021.
16. Publications Office of the European Union. European Data Portal. https://data.europa.eu/en, Last accessed: 2022-05-27.
17. Publications Office of the European Union. European Data Portal SPARQL Endpoint. https://data.europa.eu/sparql, Last accessed: 2022-05-30.
18. Publications Office of the European Union. *Open data maturity report 2021*. Publications Office, LU, 2022.
19. Luigi Reggi and Sharon S Dawes. Creating Open Government Data ecosystems: Network relations among governments, user communities, NGOs and the media. *Government Information Quarterly*, page 101675, 2022.

20. Anthony Simonofski, Anneke Zuiderwijk, Antoine Clarinval, and Wafa Hammedi. Tailoring open government data portals for lay citizens: A gamification theory approach. *International Journal of Information Management*, 65:102511, 2022.
21. Twitter, Inc. Twitter API v2. https://developer.twitter.com/en/docs/api-reference-index#twitter-api-v2, Last accessed: 2022-05-30.
22. Bastiaan Van Loenen, Anneke Zuiderwijk, Glenn Vancauwenberghe, Francisco J Lopez-Pellicer, Ingrid Mulder, Charalampos Alexopoulos, Rikke Magnussen, Mubashrah Saddiqa, Melanie Dulong de Rosnay, Joep Crompvoets, et al. Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda. *JeDEM-eJournal of eDemocracy and Open Government*, 13(2):1–27, 2021.
23. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016.
24. Xiaohua Zhu and Mark Antony Freeman. An evaluation of US municipal open data portals: A user interaction framework. *Journal of the Association for Information Science and Technology*, 70(1):27–37, 2019.