IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 May 2024, accepted 6 June 2024, date of publication 13 June 2024, date of current version 21 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3414282

RESEARCH ARTICLE

Identifying the Evolution of Open Government Data Initiatives and Their User Engagement

ABDUL AZIZ, (Member, IEEE), DAGOBERTO JOSÉ HERRERA-MURILLO, JAVIER NOGUERAS-ISO, JAVIER LACASTA, AND FRANCISCO J. LOPEZ-PELLICER

Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain

Corresponding author: Abdul Aziz (abdul.aziz@unizar.es)

This work was partially supported by the Aragon Regional Government through the project T59_23R. The work of Abdul Aziz and Dagoberto José Herrera-Murillo is supported by the ODECO (Towards a sustainable Open Data ECOsystem) 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.

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 or web forms, and some of the user experiences are reported directly in the 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 *X* (formerly Twitter): a forum used by all types of stakeholders and publicly available for consistent analysis. This work proposes a methodology for analysing the evolution of Open Government Data initiatives and their user engagement along a temporal period. First, a set of variables are collected to describe the main features of Open Data initiatives and their associated social network activity. Then, to analyse these collected data from a multidimensional and temporal perspective, we apply the well-known technique of self-organizing maps to find hidden correlations between the status of different initiatives in the analysed period. Finally, as the number of map nodes is still too big to identify clear levels of maturity, a clustering algorithm is applied to group initiatives with a similar evolution status. The feasibility of this methodology has been tested by analysing 27 European Open Government Data portals between 2017 and 2021.

INDEX TERMS Open government data, open data portals, metadata quality, user engagement, social media.

I. INTRODUCTION

In the current digital world, the notion of open data as a strategic form of public knowledge has been developed. The primary argument is that the availability of open data may serve as a catalyst for social innovation and citizen empowerment, which is closely related to the rise of open government. However, the status of open data use and involvement raises concerns [1]. The Open Data movement is spreading at a rapid pace, and the ever-increasing availability of data through Open Data portals is fuelling the development of this movement. Open Data is a movement that aims to make data more accessible to the public [2], [3]. The release of this data in forms that are both open and reusable is being facilitated by the implementation of open data initiatives and the establishment of Open Data portals by governments worldwide. The idea behind Open Data openness is that users should be able to freely access, utilize, and share the data in whatever manner they want.

Open Data portals are a sort of digital library since they are online catalogues that include descriptions of datasets, known as metadata. These kinds of catalogues make it possible to find and manage metadata records describing datasets that are either accessible online or may be downloaded in a variety of distribution formats. Furthermore, metadata records facilitate the use and reuse of datasets by providing details of authorship, provenance, and license, among other details [4], [5], [6]. Indeed, the use and reuse of data from the public sector is a crucial aspect that is driving the present trend of opening up government data [7], [8]. Open Government Data (OGD)

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

portals play a critical role in opening the data, and the constant publishing of open data in OGD portals increases the demand for data of a high quality as well as a higher quality in the portal itself. However, the majority of existing open data ecosystems are not user-driven and thus fail to properly balance supply and demand. Although the importance of users in shaping open data ecosystems is well acknowledged, existing ecosystems are mostly influenced by service providers [9].

The participation of users is essential to make existing OGD initiatives more user-oriented. Some Open Data portals already offer specialized forums or online forms where diverse groups of users may report on their experiences reusing the datasets available on these portals. Other initiatives even provide users access with specialized tools for storytelling to narrate their experiences with OGD datasets [10]. However, these feedback mechanisms are, in general, very heterogeneous and the input obtained from users is rarely accessible by the general public to be compared across different OGD initiatives. Given this lack of matured feedback mechanisms, this paper proposes to employ social networks as one of the main sources to investigate user involvement in OGD initiatives. Social networks function as an open forum in which a variety of stakeholders may share their perspectives about any kind of activity or organization. In addition, social media platforms have the potential to enhance the visibility by driving visitors, engaging them via the presentation of data and portal functions, and motivating them to return [11]. With respect to the selection of the social network that better depicts the involvement of users in Open Data portals, *X* (formerly Twitter) seems to be one of the most practical sources. Apart from being used to discuss subjects ranging from personal to professional interests, there is a growing trend to share academic content and knowledge [12]. Furthermore, according to studies performed within the context of European Union [11], *X* is the most extensively used social media channel for OGD initiatives.

The purpose of this study is to propose a methodology for analysing the evolution of OGD initiatives, and in particular, their user engagement. As a first step of the methodology, we propose to define a set of variables compiled along a time-period frame that characterize both the main features of the Open Data initiatives and the activity related to these initiatives that has been reported in the *X* social network. Then, to analyse the situation of OGD initiatives from a multidimensional and temporal perspective, we propose a combined use of self-organizing maps (SOM) and clustering techniques. On the one hand, SOM allows the distribution of OGD initiatives over a two-dimensional map with a reduced number of nodes. Each node represents a neuron of the SOM neural network, which has identified hidden partial correlations among the data, characterizing the initiatives classified within this node for a particular date. On the other hand, we propose to apply an agglomerative clustering algorithm over SOM neurons to identified uniform areas in the SOM map, which represent initiatives with a similar status of development and user engagement. The classification of initiatives into different clusters in the analysed time period allow us to establish trajectories of development and detect which types of initiatives are more prone to evolve into a more matured status. This methodology is an extension of an initial version of the same authors [13] which focused on identifying different groups of initiatives at a particular date by applying a simple factor analysis of the variables and deriving clusters in terms of the identified factors. In contrast to this initial approach, the methodology proposed in this paper considers the temporal evolution of these initiatives and the normalization of variables with respect to the size of the country. The feasibility of our proposed methodology has been tested by conducting an in-depth study of 27 European OGD portals during the period of 2017 to 2021, collecting variable data at a yearly rate.

The rest of the paper is structured as follows. Section II 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 III. Section IV presents the findings and results of this research. Section V provides insights about our findings with respect to the Open Data Maturity Reports. We conclude with a summary of the contributions and some ideas for future work.

II. RELATED RESEARCH

There are several research works in the literature that have proposed frameworks for monitoring the quality of Open Data portals. For instance, Kubler et al. [4] proposed a framework of 21 metrics to evaluate the metadata of Open Data portals in five quality dimensions: existence of properties describing key aspects of datasets such as the access, discovery, contact, rights, preservation, or temporal/spatial coverage; conformance of the content of some properties (e.g., URLs, e-mail, formats); retrievability of datasets and resources; accuracy of format and file size; and an Open Data dimension assuring the existence of open and machine readable formats. Nogueras-Iso et al. [14] proposed a framework consisting of different quality controls on Open Data Metadata with quality elements and measures inspired by the ISO 19157 standard for geographic information quality. Apart from completeness and consistency, their approach reviews exhaustively the correctness of temporal, positional, and attribute information. After testing this approach on the Spanish OGD initiative, the quality indicators revealed that accuracy and correctness of metadata should be improved. Furthermore, Máchová and Lněnička [15] also proposed a framework to assess the quality of Open Data portals on a nationwide basis in the Czech Republic. Their results indicate that there is a need for quality standards and that Open Data portals differ in the number of provided datasets as well as in the level of sophistication of the offered services. More focused on transparency aspects, Lourenço [16] proposed a set of criteria that Open Data portals should meet. This work concludes that entity coverage, information types, information seeking strategies, and data quality features are significant factors to ensure transparency

and accountability. In addition, there are also works that have investigated the influence of transparency as a design concept for Open Data portals [17], [18]. By correlating certain literary features with various phases of the transparency cycle, Open Data portals should be able to fulfil the transparency requirements.

The previous works are relevant to have an overall perspective of the current status of Open Data initiatives, their maturity or their commitment to FAIR principles [19], [20]. However, they do not take into account any insights of the direct opinion of user engagement. Moving forward to the analysis of the user perspective with respect to the interaction with prevalent Open Data platforms (e.g., CKAN, DKAN, Socrata etc), there are several works that have examined the technical commons, approaches, features, and methodologies provided by each platform, as well as their visualizations tools [21], [22]. In addition, they explored the question of why these platforms are significant to users like providers, curators, and end-users, as well as the question of what the most important publishing alternatives are accessible on these platforms.

With a higher emphasis on the analysis of user interactions in Open Data portals, Begany and Gil-Garcia [23] 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. Relying on a more manual and qualitative approach, Nikiforova and McBride [24] proposed a survey to analyse and compare the various contexts regarding the employment of OGD portals by users and emphasising the most often disregarded user-centred aspects. This work has resulted in the validation of a paradigm for the usability analysis of OGD portals, as well as the identification of the strengths and flaws of portal usability that are similar across settings. In the same line, Zhu and Freeman [25] evaluated various approaches to user interactions with OGD Initiatives. They developed a user interaction framework, in which they evaluated the United States Municipal Open Data portals and provided the findings regarding user understanding and engagement with the data portals.

Concerning the evaluation of Open Data portals in the context of the European Union, it is worth noting the existence of the Open Data Maturity Report released by the Publications Office of the European Union [11] on a yearly basis. This report mentions four dimensions for the analysis of initiatives: policy, impact, portal, and quality. In the portal dimension, it includes a sustainability variable that identifies actions applied to promote the visibility of the portal, including social media presence. According to this report, *X* (formerly Twitter) is the most widely used social media channel for this purpose.

Although there are numerous works using social media as the main source for investigating the impact of public and private organizations [26], [27], the influence of users [28], or the dissemination of scientific publications [29], there are relatively few works using social media for studying the impact of Open Data portals on the user community. Most of the existing works focus on the dissemination of datasets. For instance, Khan et al. [30] explored data citation and reuse practices in 43,802 openly available biodiversity datasets. The altmetrics sourced from blogs, *X*, 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. Likewise, Hou et al. [31] conducted a study that investigates the distribution of datasets on *X* among academics and the general public. After an analysis of 2,464 datasets from Altmetric.com, they identified viral and diverse dispersion patterns within one or two diffusion levels in social networks. Last, we must mention our previous work for analysing user involvement in Open Data initiatives [13] which focused on identifying different groups of initiatives at a single date by applying a simple factor analysis of the variables and deriving clusters in terms of the identified factors.

The approach proposed in this paper for identifying the evolution of OGD initiatives incorporates ideas from above cited approaches. On the one hand, the quality indicators proposed in works about the monitoring of Open Data portals can be used as potential variables to be considered in our multidimensional and temporal analysis. On the other hand, the works investigating the impact on social media provide alternative approaches and indicators that should be considered to measure the activity on social networks like *X*. Finally, our previous work on the analysis of Open Data portals demonstrates the interest of using factor analysis, clustering techniques and other related techniques like self-organising maps to analyse multidimensional problems [32], [33], [34].

III. METHODOLOGY

This study takes a quantitative approach to the analysis of a variety of indicators about the Open Data portals that are maintained at national level by EU member nations. As indicated in Figure 1, our proposed research approach consists of five stages: selection of portals; selection of variables characterizing each portal; collection of data for each variable; application of the SOM technique to reduce the dimensionality of variables; and the clustering of SOM results. These stages are described in the following subsections.

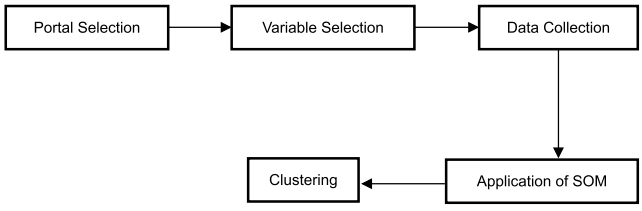


FIGURE 1. Proposed methodology for data processing.

A. PORTAL SELECTION

The portal operates as the primary source from where important data is obtained for our research analysis. So, prior to employing this method, it is necessary to locate pertinent resources in order to pinpoint the locations of portals and the papers that characterize their capabilities. The first source is the list of national catalogues in the European Data Portal [35]. The second source is the compilation of the data portals done by Juana-Espinosa and Lujan-Mora [2] where the list of platforms that were researched can be found. Both of these sources have a high level of agreement with each other.

B. VARIABLE SELECTION

This refers to the process of selecting variables that adequately characterize the properties of the Open Data portal for inclusion in our methodology. Table 1 displays the pertinent variables together with the sources from which they can be derived. The selection of variables was driven by both theoretical considerations and data availability. From a theoretical perspective, we have replicated the methodological approach followed by other authors [2], [36] who have benchmarked open data initiatives using a combination of technical indicators about portal performance and few other metrics related to impact dimensions that have not already been covered in current literature. As a result, some of the most representative operational characteristics are taken from the European Data Portal (Number of datasets - ND, Open Data Maturity score - ODM). These variables are commonly used in other similar studies [2], [36]. For introducing the new social media impact dimension, we included metrics of social activity on X (Number of Tweets - NT, Tweets from Portal - TFP, User Tweets - UT, and Number of Interactions - NI). Additionally, we added a metric of academic impact for exploratory purposes (Google Scholar - GS).

TABLE 1. Description of the variables collected for each OGD initiative and year.

Variable	Description	Source
ND	Number of datasets available for consultation / log of population	European Data Portal SPARQL API and statistics REST API
ODM	Open Data Maturity score (0-100)	Reports at European Data Portal
GS	Number of items in Google Scholar citing the portal / log of population	Google Scholar
NT	Number of relevant Tweets / log of population	*X* API
TFP	Number of Tweets from portal account / log of population	*X* API
UT	Number of users posting Tweets / log of population	*X* API
NI	Number of interactions generated by Tweets (sum of retweets, replies, quotes and likes) / log of population	*X* API

In addition, it must be noted that the raw value of some of the selected variables (e.g. number of datasets or number of users in social networks) is clearly proportional to the size of the country behind the OGD initiative. Therefore, we decided to normalize the values of these variables dividing by the log of the population of the country at each analysed year. The only exception is the ODM variable, as this refers to a qualitative measure of maturity reported by experts that take into consideration the whole context of the initiative.

C. DATA COLLECTION

This step refers to the process of collecting data of the selected variables for each OGD initiative and year in the analysed period. Regarding the variables that were gathered manually, it is important to point out that the ODM variable was extracted from the reports published by the European Commission [11]. In the case of GS variable, a manual search in Google Scholar for the number of publications citing the homepage URL of each OGD initiative was carried out. In addition, it must be noted that this search was performed for each year in the analysed period by adding a temporal filter on the citing publications.

With respect to the values collected automatically, the collection of values associated with the ND variable was not an easy task. Although the European Data Portal facilitates an SPARQL endpoint to query the Open Data collected from the different national initiatives [37], the temporal information contained in metadata records is not a completely reliable source. The *dcat:Dataset* entity of metadata records (compliant with the DCAT-AP vocabulary) includes *dcat:created*, *dct:modified* and *dct:issued* properties to inform about the creation date, the modification date and the publication date of a dataset. However, either this information is sometimes missing or it does not explicitly imply that a dataset was directly published in a national data portal. A most reliable source also available at the European Data Portal are the statistics compiled as a result of the harvesting processes performed along time from the different catalogues of the OGD initiatives.[1] This statistical information can be queried through a specific API.[2] The problem is that this information is only available from 2019 onwards. Therefore, in order to estimate the missing number of datasets for the years 2017 and 2018 for each portal, we assumed a constant annual growth rate for the period with the available data between 2019 to 2022 and projected that growth rate to the previous two years where there is no data. We computed this constant annual growth rate, named r, for each portal using the following equation:

$$datasets\_at\_2019 \cdot (1 + r)^3 = datasets\_at\_2022$$

Consequently, the formula for obtaining r is as follows:

$$r = \sqrt[3]{\frac{datasets\_at\_2022}{datasets\_at\_2019}} - 1$$

[1] https://data.europa.eu/catalogue-statistics/evolution/countryCatalogue?locale=en

[2] https://data.europa.eu/api/hub/statistics/data/ds-per-catalogue?list=true

With respect to the variables that measure the online social network activities on *X* connected to portals from 2017 through 2021 (NT, TFP, UT, and NI), they were gathered at the end of the year 2022 with the use of the *X* API for Academic Research [38]. This API makes it possible to get tweets whose content either references the URL of portals or their *X* accounts.

D. APPLICATION OF THE SELF-ORGANIZING MAP (SOM) TECHNIQUE

A Self-Organizing Map (SOM) is an artificial neural network method that performs dimensionality reduction over an input dataset. The standard SOM algorithm involves an unsupervised neural network with competitive learning and no hidden layers [39]. The objective is to align an input vector (representation of the variables describing an element of the dataset) with a neuron in an output matrix of neurons. The SOM maintains the topology of input characteristics while simultaneously reducing the number of dimensions in a dataset and making this dataset easier to understand [34].

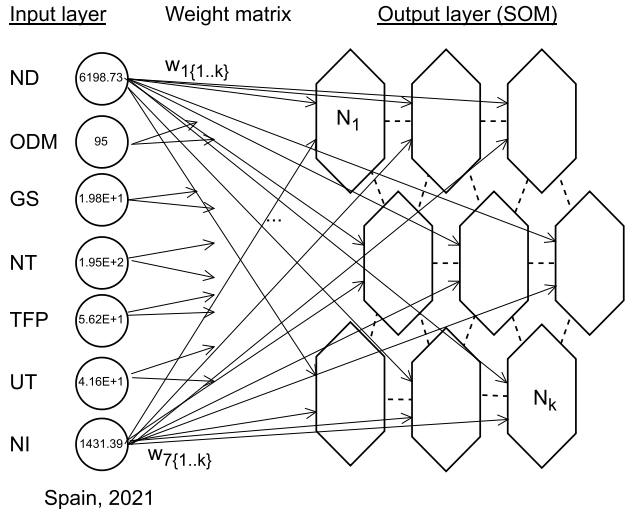


FIGURE 2. An example of a small SOM trained with a dataset consisting of records describing OGD initiatives.

Figure 2 presents an example of a SOM trained for improving the visualization in a lower number of dimensions of a dataset consisting of the records describing an OGD initiative at a particular year with the variables enumerated in section III-C. During the training of this neural network, each of the vector components in the input layer is connected to the complete output map (output layer) by a weight matrix. It can be observed in the example that we have chosen a hexagonal topology, i.e. each central neuron has 6 neighbours. In contrast to other networks, the surrounding neurons of the output topology have an impact on the weights when the network is being trained. In addition, the output map in the example has a size of 9 ($k = 9 = 3 \times 3$). The size of the network must be set up during the experimental phase in order to accommodate an appropriate number of neurons in x and y directions according to the original size of the dataset.

E. CLUSTERING

The use of a SOM model allows a clearer analysis in terms of the linkages between the produced aggregations. However, as the number of nodes in the SOM map is too big to identify similar levels in the status of development of OGD initiatives, this phase of the methodology proposes to apply a clustering to the nodes of the SOM map. Data points (i.e. initiatives at different years) that are similar to one another are grouped together in clusters according to the underlying patterns and connections that they share.

We propose to use an agglomerative clustering algorithm [40], i.e. a hierarchical clustering technique that follows a bottom-up approach to partition the data generating a hierarchical structure progressively. In particular, we use use the Ward clustering algorithm [41].

IV. EXPERIMENTS AND RESULTS

This section presents the outcomes of applying the proposed methodology on the national Open Data portals of 27 EU member countries and their online social network activities on *X* during the temporal range from year 2017 to year 2021. This represents an input dataset consisting of 135 records: each record describes the status of the 27 national initiatives at each of the analysed years. It must be noted that Hungary, also belonging to the European Union, has not been considered in this study because there is not an official open data portal.

Figure 3 presents an overview of the input records and the values contained in the 7 considered variables over a bi-dimensional space using principal component analysis (PCA). This figure helps to guess the clouds of points that could be the origin of the clusters that are later identified. PCA is a method for reducing the number of dimensions that is often used to convert complicated data into a space with fewer dimensions while maintaining the fundamental variance of data [42]. In can be observed that there are 3 separate clouds of points in this graphical representation: a small cloud of points on the left upper corner; a small cloud of points on the right side; and a bigger cloud of points on the left side.

As indicated in the description of our methodology, the core tool for the analysis of input data is the generation of a SOM map combined with the clustering of map nodes. The first parameter to be decided for the generation of a SOM map is its dimension, i.e. the number of rows and columns in this map. According to the recommendations of Vesanto and Alhomieni [43], we decided to approximate the number of neurons (cells) in the map to $5 \times \sqrt{data\ input\ size}$. Therefore, we selected 8×7 neurons, i.e. 8 rows and 7 columns in a bi-dimensional map. Figure 4 shows the SOM (8×7 dimensions) map obtained after training the SOM neural network with 100,000 steps (epochs). This figure also helps in the identification of the 4 clusters grouping the nodes. The SOM map nodes represent the classification of records (initiatives with a particular status at each year) in the different output neurons. The colour gradation in these nodes provides

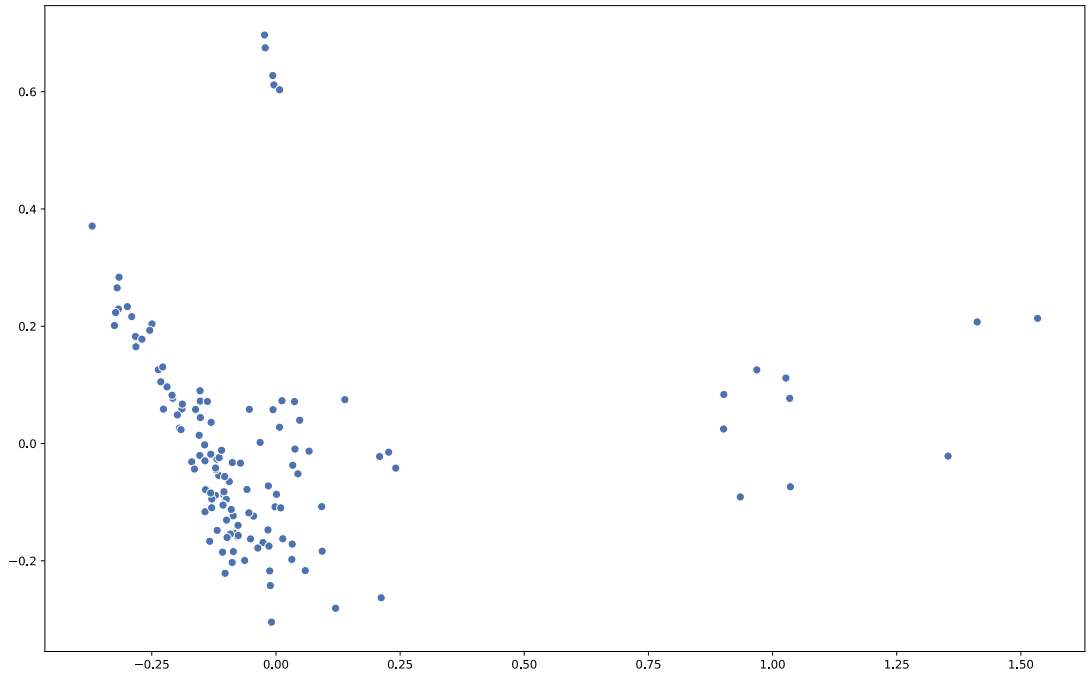


FIGURE 3. Dispersion of records in input dataset after applying PCA over a bi-dimensional space.

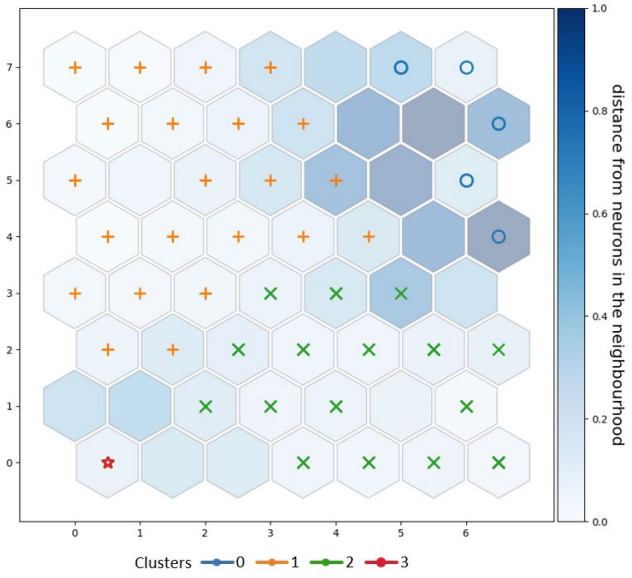


FIGURE 4. Output SOM with 4 clusters and 8 × 7 dimension.

clusters that identify similar levels of development in Open Data initiatives we also generated a dendrogram (see Figure 5). A dendrogram is a complementary output to verify that the selected number of clusters groups appropriately the records of the input.

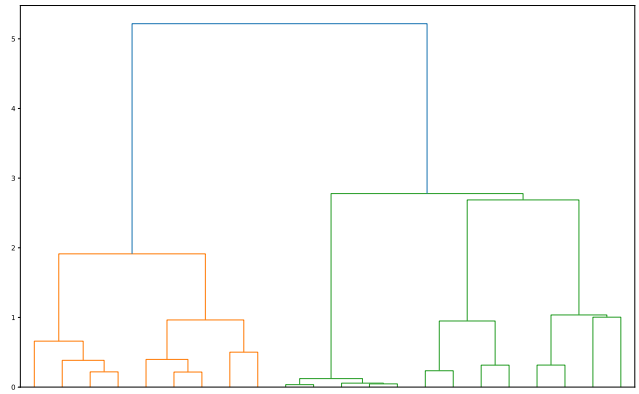


FIGURE 5. Cluster dendrogram.

an indication of the distance of variable values with respect to the neurons in the neighbourhood.

The map in Figure 4 also displays the classification of the neurons (node maps) in four clusters after applying an agglomerative clustering algorithm. We performed our experiments as exploratory analysis on the different values of k (number of clusters) to see the optimal result. We tested from $k = 2$ until $k = 7$ and we found that with $k = 4$ we get the optimal results for this study. Hence, we chose $k = 4$ as the number of clusters. For a better selection of the number of

In order to have a better understanding of the meaning of these clusters, Figure 6 provides a cluster profiling plot with the average normalized score of the seven considered variables for each cluster which is represented by a line. Some details of these clusters are as follows:

- Cluster 0 has the highest values for most of the variables. This cluster contains the largest proportions of number of tweets (NT), user tweets (UT), and tweets from the portal (TFP), showing that this category sees a substantial level of X activity overall. In addition to this, it stands out in

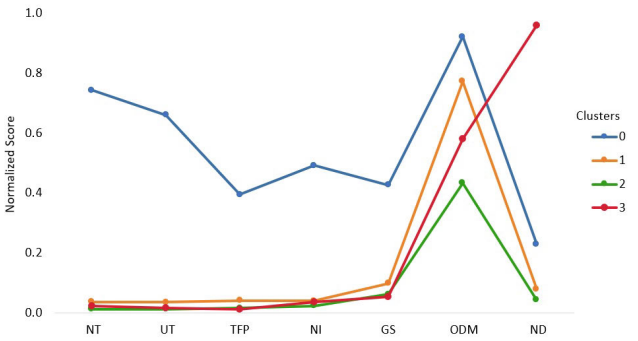


FIGURE 6. Profiling of clusters.

terms of the number of interactions (NI), which suggests that the information included inside this cluster produces a significant amount of engagement. The comparatively high values for Google Scholar (GS) mentions, Open Data Maturity (ODM) Score, and number of datasets (ND) all point to the fact that this cluster is academically acknowledged, advanced in terms of open data policies, and abundant in datasets that are readily accessible.

- Cluster 1, on the other hand, appears to reflect lower values than cluster 0 across the board of all parameters. This cluster shows a decrease in the number of tweets (NT), user engagements (UT), and portal activity (TFP). In addition, the number of Google Scholar (GS) mentions, the Open Data Maturity (ODM) Score, and the number of datasets (ND) associated with this cluster are all much fewer than those associated with the initiatives classified in cluster 0. Based on this information, cluster 1 seems to have less impact, a less developed set of open data standards, and maybe fewer datasets available.
- Clusters 2 and 3 are situated below clusters 0 and 1 for most of the considered variables:
  - Cluster 2 reveals a moderate presence of tweets (NT), user interaction (UT), and portal activity (TFP), in addition to relatively modest values for Google Scholar (GS) mentions, Open Data Maturity (ODM) Score, and number of datasets (ND). This cluster also reveals a moderate presence of open datasets.
  - Cluster 3 provides somewhat higher statistics across all parameters compared to cluster 2, showing a better degree of activity, recognition, maturity, and dataset availability. Its relevance within the existing context is shown by the fact that the even number of datasets is greater than that of the other three clusters.

In addition, for each variable we conducted an ANOVA test to detect the presence of differences between clusters. In the cases where this test was significant (p <0.05), we accompanied it with post hoc tests to compare the clusters and detect the precise origin of the differences. We observed that for the seven variables the ANOVA test was significant, which means that at least one of the clusters is statistically different from the rest. For the NT, UT, TFP, NI and GS variables, the post hoc test revealed that the cluster 0 is statistically different from the rest and there is no difference between the rest of the clusters. For the ODM and ND variables, all the comparisons were significant which means that all the clusters have significantly different values.

Making a deeper analysis of the countries behind the initiatives and years grouped in the different nodes and clusters displayed in the map shown in Figure 4, it can be derived the following:

- The red star in cluster 3 represents the open data portal for the Czech Republic. Based on the findings of our experiment, this is an anomaly or an outlier because the number of datasets (ND) published each year in the Czech Republic open data portal are significantly higher than in the other EU Open Data portals.
- The Open Data portals of Spain and France are clustered together in the blue circles of cluster 0, primarily due to their consistently high values across various selected variables. France and Spain stand out with the highest nominal values for key aspects, such as Twitter discussions (NT, TFP, UT, and NI) and Google Scholar (GS) mentions, reflecting their greater significance within the open data portal landscape.
- Meanwhile, the development of other Open Data portals is represented with orange plus signs in cluster 1 and green crosses in cluster 2.

The distribution of records representing the initiatives at different years can be also observed in Figure 7, which is equivalent to Figure 3 displaying the records in a bi-dimensional space using PCA. The novelty of Figure 7 is that we highlight now the assignment of records (points in the plot) to the four clusters obtained after applying the clustering algorithm on map nodes. The records corresponding to the Czech Republic in cluster 3 are the points on the left-upper corner. The points corresponding to the records of Spain and France in cluster 0 are the ones on the right side. The points corresponding to clusters 1 and 2 are on the left side. The points of cluster 1 corresponding to initiatives with a relatively more mature status are closer to the horizontal axis.

Another advantage of using this approach for the analysis of the evolution of OGD initiatives is the possibility of analysing the trajectories or movements of initiatives along the clusters in the studied time period. Table 2 shows the evolution of each country's open data portal along the 5-year time frame period. Some observations can be highlighted about countries remaining or moving between the clusters:

- With respect to countries remaining on the same cluster, it is worth noting that France and Spain have stayed in cluster 0 from 2017 to 2021. This is due to the fact that both of these countries are doing extremely well in the field of open data and in open data society. In a similar fashion, Austria, Cyprus, Ireland, Italy, the Netherlands, and Slovenia have all remained in cluster 1 over the whole of the time period covered by the experiment in this research study. In addition, given the values of the

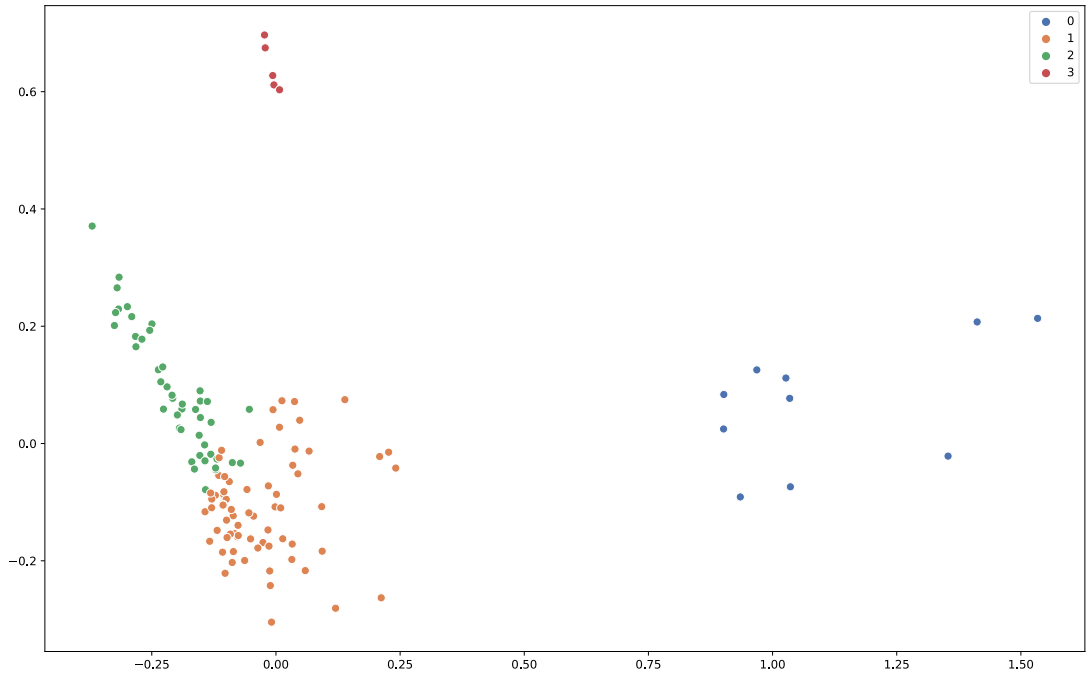


FIGURE 7. Dispersion of records in input dataset after applying PCA over a bi-dimensional space, with assigned cluster.

TABLE 2. Cluster shifting of the countries.

Country	2017	2018	2019	2020	2021
AUSTRIA	1	1	1	1	1
BELGIUM	2	1	2	2	2
BULGARIA	1	1	2	2	2
CROATIA	1	2	1	1	1
CYPRUS	1	1	1	1	1
CZECH	3	3	3	3	3
DENMARK	2	2	1	1	1
ESTONIA	2	2	2	1	1
FINLAND	1	2	1	1	1
FRANCE	0	0	0	0	0
GERMANY	2	1	1	1	1
GREECE	1	1	2	1	1
IRELAND	1	1	1	1	1
ITALY	1	1	1	1	1
LATVIA	2	1	1	1	2
LITHUANIA	2	2	2	1	1
LUXEMBOURG	1	1	2	2	2
MALTA	2	2	2	2	2
NETHERLANDS	1	1	1	1	1
POLAND	2	1	1	1	1
PORTUGAL	2	2	2	2	2
ROMANIA	1	2	2	2	2
SLOVAKIA	1	1	2	2	2
SLOVENIA	1	1	1	1	1
SPAIN	0	0	0	0	0
SWEDEN	2	2	2	1	1

chosen variables that were taken into consideration for the purposes of this study, the open data portal for the Czech Republic is an outlier that has not shifted from cluster 3 over the temporal range of years from 2017 to 2021. This is because the open data portal for the Czech Republic is the only one that publishes a higher number of datasets (ND) than any other country, making it an outlier, as can be clearly seen from Figure 6. Although this cluster contains only one country, we decided to include it because it is representative of a strategy that prioritises the quantity of datasets over other variables.

- On the other hand, there have been some movements between cluster 1 and cluster 2, or vice versa. Denmark, Estonia, Germany or Poland started in cluster 2 and moved to cluster 1 in the following years. This can be interpreted as an improvement in the maturity status of their initiatives (value of ODM score), as the values of variables in cluster 1 are higher than those in cluster 2. On the contrary, Bulgaria, Luxembourg, Romania and Slovakia started in cluster 1 and move to cluster 2, which can be interpreted as a decrease in their maturity status.
- Last, several national Open Data portals have moved from an initial cluster to another cluster, and then returned back to the initial cluster. For instance, countries like Belgium started out in cluster 2 in the year 2017, moved up to cluster 1 in the year 2018, and finally returned back to cluster 2 for the next three years. The same thing happened to countries like Croatia, but with a different cluster: Croatia started out in cluster 1 in 2017, moved up to cluster 2 in 2018, and then shifted back to cluster 1 for the next three years.

The full code of the experiments is available on a GitHub repository.[3] We employed external Python libraries as well as functions that we developed ourselves.

[3]Github Repository https://github.com/IAAA-Lab/Evolution-of-OGD-Initiatives.

V. DISCUSSION

The feasibility of our analysis methodology was tested by evaluating the development of 27 European OGD portals between 2017 and 2021. Using as input the values of the selected variables at a yearly basis, we were able to compare the output of our methodology with the conclusions reported in the Open Data Maturity Report for the years ranging from 2017 to 2021.

In the course of our experiment, we used a SOM-based model and a clustering algorithm in order to identify different maturity levels in the evolution of OGD initiatives according to the variables that were selected. Our first observation is that cluster 0 in Figure 6 could be considered as ''user community driven'' because the initiatives in this cluster have the highest values for many variables and the highest concentrations of Number of Tweets (NT), User Tweets (UT), and Tweets from the portal (TFP), all of which indicate a high volume of *X* activity. Not only does the content within this cluster generate a large number of interactions (NI), but it also stands out in terms of engagement volume. In addition, it is highly interdependent. The relatively high values for Google Scholar (GS) mentions, Open Data Maturity (ODM) Score, and Number of Datasets (ND) indicate that this cluster is well-known among academics, mature in terms of open data policy, and abundant in freely accessible datasets. Cluster 0 Open Data portals are in the greatest position overall and may be defined as user-centric, consistent with the findings of the 2017-2021 Open Data Maturity Report. This may be largely attributable to the activities of the Open Data portals in cluster 0. These initiatives monitor user feedback through multiple channels, such as message forums dedicated to each dataset (e.g., in case of the French open data portal). The Open Data Maturity Reports of 2017-2021 highlight the efforts of cluster 0 Open Data portals to cultivate editorial content, improve search and findability, and make active use of social media platforms such as Facebook, LinkedIn, YouTube, and Flickr. In addition, from Figure 6 it can be observed that cluster 3 includes open data initiatives with a remarkable number and quantity of released datasets. However, these activities have no direct effect on *X* activity. This is likely due to the fact that the bodies responsible for coordinating these efforts are not as effective as they could be at promoting their work on social media platforms. Lastly, we are able to determine that the open data initiatives in clusters 1 and 2 report a moderate level of quality and social network activity. This most likely indicates that the initiatives in question are improving and in their infancy, making them less appealing to potential consumers.

In addition, it is worth noting the work involved in collecting the values of the various variables that were special for this research. The information may be used by governments and decision-makers in order to assess Open Data initiatives based on the level of user participation that they have received. In particular, the variables on X activity indicators (NT, UT, and NI) and Google Scholar mentions (GS) were specifically chosen to capture the user engagement. Moreover, it must be noted that to increase the applicability of our experiment and generate a better distribution of initiatives in the SOM map, we decided to normalize the raw values of the variables dividing them by the logarithm of the population of each country in each of the analysed years. This allowed us to decrease the effect of comparing initiatives in countries with big differences in terms of population and governmental complexity. The only exception was the ODM variable, as this indicator is a score assigned manually by experts taking into account the overall context of each initiative and country. During the development and testing of our methodology, we also tested the inclusion of a variety of other variables and their permutations, such as the sentiment score of tweets from each country over the span of each year, both individually and collectively, and the number of likes from each year in each country and number of positive tweets of each country for our temporal range of years from 2017 to 2021. However, as a consequence of their negative impact on the generation of the SOM map and the clustering phase, certain variables and combinations were eliminated.

Last, although our research provides some useful insights, it also has a few drawbacks. First, the experiments have been performed on data that is open to the public, and the reliability of the results is dependent on the quality of the data sources. In addition, the research is limited to a certain time period (2017-2021), and it is possible that the dynamics of OGD portals have progressed since then. In addition, our method of clustering is predicated on a number of different characteristics. Nevertheless, there may be more important variables that have an impact on the level of user participation and data compliance.

VI. CONCLUSION

In this paper, a methodology for conducting an in-depth study of 27 European Open Government Data portals during the period of 2017 to 2021 has been presented. The study concentrates on the portal activity in (one of the most widely used social networks) and open data maturity levels. We acknowledge that there are some OGD initiatives offering discussion forums, chats and other feedback mechanisms, but, as stated in the introduction, these ad-hoc feedback mechanisms do not have a standardized way to retrieve the input provided by users. Through the use of our methodology, we were able to monitor the development of each portal in relation to its Open Data Maturity Report during this time period. Additionally, we made an effort to comprehend the connection that exists between the activity on social networks and the primary characteristics that define the size, quality, and level of development of open data initiatives.

Moreover, pertinent inferences are able to be formed from the analysis obtained via the use of clustering methods and Self-Organizing Maps (SOM). SOM reduces the high-dimensions of input data into a format that is visually accessible, and together with the use of clustering techniques, has proven to be a useful tool for understanding and monitoring the development of OGD initiatives and the user interaction associated with them. We were able to

cluster OGD initiatives, identify patterns of user involvement, and recognize shifts and changes through time. However, it is evident that the proposed methodology also requires a manual effort for the interpretation of the computational results. In order to transform the results into useful insights, we need a qualitative reasoning about the formation and movements between clusters. This manual interpretation is necessary to provide stakeholders in charge of the development of Open Government Data portals with the necessary advice for improving user engagement and increasing openness and accountability in governance.

There are also several potential lines for further study. The quality of portals might be enhanced by gaining a deeper insight of the factors that contribute to varying degrees of user engagement and data conformity. On the one hand, although we already observed that the inclusion of a variable considering the sentiment analysis of tweets did not provide an impact on the construction of clusters, we believe that the categorization of tweet contents remains a promising area for future study. This approach can have the capacity to provide more valuable and enlightening information on specific aspects of user engagement (e.g., data quality concerns, suggestions for improvement, data reuse examples). On the other hand, we can extend the experiments to OGD initiatives outside the European context and also consider a longer time period to analyse the evolution of OGD initiatives. Finally, policymakers and practitioners may benefit from better OGD initiatives if researchers compare data across regions or countries to identify regional differences and best practices.

REFERENCES

[1] G. Santos-Hermosa, A. Quarati, E. Loría-Soriano, and J. E. Raffaghelli, "Why does open data get underused? A focus on the role of (open) data literacy," in *Higher Education Dynamics*. Cham, Switzerland: Springer, 2023, pp. 145–177.

[2] S. de Juana-Espinosa and S. Luján-Mora, "Open government data portals in the European union: A dataset from 2015 to 2017," *Data Brief*, vol. 29, Apr. 2020, Art. no. 105156.

[3] L. Reggi and S. S. Dawes, "Creating open government data ecosystems: Network relations among governments, user communities, NGOs and the media," *Government Inf. Quart.*, vol. 39, no. 2, Apr. 2022, Art. no. 101675.

[4] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, and Y. Le Traon, "Comparison of metadata quality in open data portals using the analytic hierarchy process," *Government Inf. Quart.*, vol. 35, no. 1, pp. 13–29, Jan. 2018.

[5] S. Neumaier, J. Umbrich, and A. Polleres, "Automated quality assessment of metadata across open data portals," *J. Data Inf. Qual.*, vol. 8, no. 1, pp. 1–29, Nov. 2016.

[6] J. Riley. (2017). *Understanding Metadata*. [Online]. Available: https//www.niso.org/publications/press/UnderstandingMetadata.pdf

[7] W. Carrara, W.-S. Chan, S. Fischer, and E. van-Steenbergen. *Creating Value Through Open Data: Study on the Impact of Re-Use of Public Data Resources European Commission, Directorate General for Communications Networks, Content and Technology*. Accessed: 2015. [Online]. Available: https://data.europa.eu/doi/10.2759/328101

[8] A. Simonofski, A. Zuiderwijk, A. Clarinval, and W. Hammedi, "Tailoring open government data portals for lay citizens: A gamification theory approach," *Int. J. Inf. Manage.*, vol. 65, Aug. 2022, Art. no. 102511.

[9] B. Van Loenen, A. Zuiderwijk, G. Vancauwenberghe, F. J. Lopez-Pellicer, I. Mulder, C. Alexopoulos, R. Magnussen, M. Saddiqa, M. Dulong de Rosnay, J. Crompvoets, A. Polini, B. Re, and C. Casiano Flores, "Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda," *JeDEM J. Democracy Open Government*, vol. 13, no. 2, pp. 1–27, Dec. 2021.

[10] L. Akerreta Escribano and J. 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*, vol. 27, no. 1, pp. 31–43, Jun. 2021.

[11] Publications Office Eur. Union. (2021). *Open Data Maturity Report*. [Online]. Available: https://data.europa.eu/doi/10.2830/394148

[12] S. Haustein, R. Costas, and V. Larivière, "Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0120495.

[13] D. J. Herrera-Murillo, A. Aziz, J. Nogueras-Iso, and F. J. Lopez-Pellicer, "Analysing user involvement in open government data initiatives," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, 2022, pp. 175–186.

[14] J. Nogueras-Iso, J. Lacasta, M. A. Ureña-Cámara, and F. J. Ariza-López, "Quality of metadata in open data portals," *IEEE Access*, vol. 9, pp. 60364–60382, 2021.

[15] R. Máchová and M. Lnenicka, "Evaluating the quality of open data portals on the national level," *J. Theor. Appl. Electron. commerce Res.*, vol. 12, no. 1, pp. 21–41, 2017.

[16] R. P. Lourenço, "Open government portals assessment: A transparency for accountability perspective," in *Proc. 12th IFIP WG 8.5 Int. Conf.*, 2013, pp. 62–74.

[17] M. Lnenič ka, R. Machova, J. Volejníková, V. Linhartová, R. Knezackova, and M. Hub, "Enhancing transparency through open government data: The case of data portals and their features and capabilities," *Online Inf. Rev.*, vol. 45, no. 6, pp. 1021–1038, Oct. 2021.

[18] M. Lnenicka and A. Nikiforova, "Transparency-by-design: What is the role of open data portals?" *Telematics Informat.*, vol. 61, Aug. 2021, Art. no. 101605.

[19] A. Aziz, "Technical aspects for inclusiveness across user domains in data portals," in *Proc. Workshops Doctoral Consortium*, 2023, pp. 271–280.

[20] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018.

[21] M. Ali, C. Alexopoulos, and Y. Charalabidis, "A comprehensive review of open data platforms, prevalent technologies, and functionalities," in *Proc. 15th Int. Conf. Theory Pract. Electron. Governance*, Oct. 2022, pp. 203–214.

[22] A. S. Correa, P.-O. Zander, and F. S. C. da Silva, "Investigating open data portals automatically: A methodology and some illustrations," in *Proc. 19th Annu. Int. Conf. Digit. Government Res. Governance Data Age*, May 2018, pp. 1–10.

[23] G. M. Begany and J. R. Gil-Garcia, "Understanding the actual use of open data: Levels of engagement and how they are related," *Telematics Informat.*, vol. 63, Oct. 2021, Art. no. 101673.

[24] A. Nikiforova and K. McBride, "Open government data portal usability: A user-centred usability analysis of 41 open government data portals," *Telematics Informat.*, vol. 58, May 2021, Art. no. 101539.

[25] X. Zhu and M. A. Freeman, "An evaluation of U.S. municipal open data portals: A user interaction framework," *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 1, pp. 27–37, Jan. 2019.

[26] H. Shahbaznezhad, R. Dolan, and M. Rashidirad, "The role of social media content format and platform in users' engagement behavior," *J. Interact. Marketing*, vol. 53, pp. 47–65, Feb. 2021.

[27] J. L. Alonso Berrocal, C. G. Figuerola, and Á. F. Zazo Rodríguez, "Propuesta de índice de influencia de contenidos (Influ@RT) en Twitter," *Scire, Representación Y Organización del Conocimiento*, vol. 21. no. 1, pp. 21–26, Jun. 2015.

[28] M. Zhang, D. Zhang, Y. Zhang, K. Yeager, and T. N. Fields, "An exploratory study of Twitter metrics for measuring user influence," *J. Informetrics*, vol. 17, no. 4, Nov. 2023, Art. no. 101454.

[29] F. Didegah, N. Mejlgaard, and M. P. Sørensen, "Investigating the quality of interactions and public engagement around scientific papers on Twitter," *J. Informetrics*, vol. 12, no. 3, pp. 960–971, Aug. 2018.

[30] N. Khan, M. Thelwall, and K. Kousha, "Measuring the impact of biodiversity datasets: Data reuse, citations and altmetrics," *Scientometrics*, vol. 126, no. 4, pp. 3621–3639, Apr. 2021.

[31] J. Hou, Y. Wang, Y. Zhang, and D. Wang, "How do scholars and non-scholars participate in dataset dissemination on Twitter," *J. Informetrics*, vol. 16, no. 1, Feb. 2022, Art. no. 101223.

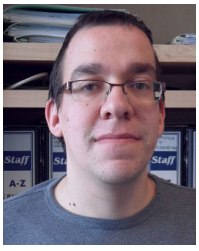
[32] B. C. Hewitson, "Climate analysis, modelling, and regional downscaling using self-organizing maps," in *Self-Organising Maps: Applications in Geographic Information Science*, P. Agarwal and A. Skupin, Eds. Hoboken, NJ, USA: Wiley, ch. 8, 2008, pp. 137–153, doi: 10.1002/9780470021699.ch8.

[33] T. Li, G. Sun, C. Yang, K. Liang, S. Ma, and L. Huang, "Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes," *Sci. Total Environ.*, vols. 628–629, pp. 1446–1459, Jul. 2018.
[34] A. Ruiz-Varona, J. Lacasta, and J. Nogueras-Iso, "Self-organizing maps to evaluate multidimensional trajectories of shrinkage in Spain," *ISPRS Int. J. Geo-Information*, vol. 11, no. 2, p. 77, Jan. 2022.
[35] Publications Office Eur. Union. *European Data Portal*. Accessed: May, 27, 2022. [Online]. Available: https://data.europa.eu/en
[36] S. de Juana-Espinosa and S. Luján-Mora, "Open government data portals in the European union: Considerations, development, and expectations," *Technol. Forecasting Social Change*, vol. 149, Dec. 2019, Art. no. 119769.
[37] Publications Office Eur. Union. *European Data Portal SPARQL Endpoint*. Accessed: May 30, 2022. [Online]. Available: https://data.europa.eu/sparql
[38] X APIV2. *Twitter API V2*. Accessed: May 30, 2022. [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index
[39] A. Skupin and P. Agarwal, "Introduction: What is a self-organizing map?" in *Self-Organising Maps: Applications in Geographic Information Science*. Hoboken, NJ, USA: Wiley, 2008, ch1, pp. 1–20. [Online]. Available: https://doi.org/10.1002/9780470021699.ch1
[40] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, Jan. 2012.
[41] V. Batagelj, *Classification and Related Methods of Data Analysis*. Amsterdam, The Netherlands: North Holland, 1988, pp. 67–74.
[42] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *J. Soft Comput. Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.
[43] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.

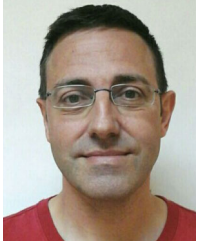
ABDUL AZIZ (Member, IEEE) received the bachelor's degree in computer science from the COMSATS Institute of Information Technology, Lahore, Pakistan, in 2013, and the master's degree in computer science from the National University of Computer and Emerging Sciences, Karachi, Pakistan, in 2018. He is currently pursuing the Ph.D. degree in computer science with the University of Zaragoza, within the Advanced Information Systems Laboratory (IAAA) of the Aragon Institute of Engineering Research (I3A). He is an Early Stage Researcher (ESR 08, "Technical aspects for inclusiveness across user domains in data portals") for the ODECO project, and Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020). His research interests include open data, information retrieval, and data portal domains.

DAGOBERTO JOSÉ HERRERA-MURILLO received the bachelor's degree in business informatics from Tecnológico de Monterrey and the dual master's degree in big data management from Université Libre de Bruxelles, BarcelonaTech, and Technische Universiteit Eindhoven. He has worked on education and volunteered on data science initiatives for social good. He is currently working towards the Ph.D. degree in computer science at the University of Zaragoza, within the Advanced Information Systems Laboratory (IAAA) of the Aragon Institute of Engineering Research (I3A). He is an Early Stage Researcher (ESR 02, "User interface design: optimising findability") for the ODECO project, an Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020).

JAVIER NOGUERAS-ISO received the M.S. and Ph.D. degrees in computer science from the University of Zaragoza, Spain. In 1998, he started his research with the Advanced Information Systems Laboratory, University of Zaragoza, where he is currently a Full Professor of computer science. From 2011 to 2017, he was the Director of the Catedra Logisman on Technological Document Management. From 2015 to 2019, he was the Associate Director of the Aragon Institute of Engineering Research (I3A). His research interests include information retrieval and semantic web technologies applied to different domains, although with a special emphasis on geographic information infrastructures.



JAVIER LACASTA received the Ph.D. degree in computer science. Since 2019, he has been an Associate Professor with the Computer Science and Systems Engineering Department, University of Zaragoza. Since 2002, he has been his research work has focused around the problem of information retrieval in geographical and bibliographical data collections. Over the last few years, he has coauthored more than 80 publications in books, journals or conference proceedings. He has also participated in several public R+D calls, and in national and European transfer contracts. His interests include geospatial data modeling, knowledge management, semantic web, information retrieval, artificial intelligence, and data mining.



FRANCISCO J. LOPEZ-PELLICER received the M.S. and Ph.D. degrees in computer engineering from the University of Zaragoza. In 2004, he started his research with the Advanced Information Systems Group (IAAA), University of Zaragoza. Currently, he is an Associate Professor with the Advanced Information Systems Group (IAAA), University of Zaragoza. Over the past ten years, his professional career has been linked to open data initiatives and spatial data infrastructures. Within this context, he has coauthored numerous publications in books, journals or conference proceedings; and has collaborated in several R+D projects. His research interests include open data infrastructures, service-based geographic information systems, and various information systems.