Open Government Data for Non-Expert Citizens

Chokki, Abiola Paterne; Simonofski, Anthony; Frénay, Benoît; Vanderose, Benoit

Publication date:
2021

Document Version
Peer reviewed version

Open Government Data for Non-Expert Citizens: Understanding Content and Visualizations' Expectations

Abiola Paterne Chokki[0000−0003−4500−2141], Anthony Simonofski[0000−0002−1816−5685], Benoît Frénay[0000−0002−7859−2750], and Benoît Vanderose[0000−0001−9752−0085]

University of Namur, Namur, Belgium
{abiola-paterne.chokki, anthony.simonofski}@unamur.be

Abstract. Open government data (OGD) refers to data made available online by governments for anyone to freely reuse. Non-expert users, however, lack the necessary technical skills and therefore face challenges when trying to exploit it. Amongst these challenges, finding useful datasets for citizens is very difficult as their expectations are not always identified. Furthermore, findings the appropriate visualization that is more understandable by citizens is also a barrier. The goal of this paper is to decrease those two entry barriers by better understanding the expectations of non-expert citizens. In order to reach that goal, we first seek to understand their content expectations through the usage statistics analysis of the OGD portal of Namur and through a complementary online survey of 43 participants. Second, we conduct interviews with 10 citizens to obtain their opinion on the appropriate and well-designed visualizations of the content they seek. The findings of this multi-method approach allow us to issue 5 recommendations for OGD portal publishers and developers to foster non-expert use of OGD.

Keywords: Open Government Data · Visualization · Content · Non-expert.

1 Introduction

Open government data (OGD) refers to data made available online by governments for anyone to freely reuse. OGD initiatives increase government transparency and accountability, but also involve many challenges such as its potential reuse by non-experts [1,2]. These challenges have been minimized by developers who have implemented many services or applications for citizens using open data. However, this approach has two issues. First, the developers and OGD publishers are not aware of the datasets that are likely to be of interest to citizens [3]. Therefore, our first research question **(RQ1)** is as follows: "What data are non-experts interested in?". Second, the OGD publishers and especially the developers are not aware of the needs of citizens in terms of appropriate visualization design to represent data in an understandable manner [4]. Our second

research question **(RQ2)** is therefore formulated as follows: "How to optimally visualize OGD to non-experts?".

The goal of this paper is to better understand citizens' expectations, the useful datasets and the adequate visualizations they expect on OGD portals. The study seeks to identify the datasets needed by non-experts by analyzing the dataset usage statistics on the open data portal of Namur, but also by doing an online survey with 43 participants to find out the real expectations of non-experts in order to provide recommendations to developers about the interested datasets to used in applications and to OGD publishers about the datasets to publish on portals. To find answers to RQ2, we conduct interviews with 10 citizens to get their opinion on the appropriate and well-designed type of visualization for OGD. The data recommendation tool NeDVis[1] under development in our laboratory is used to support the interviews.

2 Methodology

2.1 Usage statistics and survey (RQ1)

In order to address RQ1, we combined three resources. First, we used the OGD portal of the city of Namur (Belgium)[2] as use case to study the actual consultation statistics of the datasets on the portal. We chose this portal as it is the most advanced portal in Wallonia (Belgium) and access with key stakeholders of this portal was possible. This information was collected through a file sent by the OGD manager of the city to the researchers. Second, the list of High-Value Datasets[3] (HVDs) representing datasets with significant benefits to society, from the data portal of the Dutch government, was used to initiate the survey. The list was also used to verify if it matches with the real expectations of non-expert users. Third, to complement the findings from these statistics, we issued a survey[4] to the citizens of Namur asking them what datasets they expect to find on portals.

The survey was implemented using Google forms and pretested by two users to ensure all kinds of errors that are associated with survey research are reduced [5]. The survey was later shared on Facebook groups and was filled in by 43 non-experts. This low participation rate can be explained by three reasons: (i) we focused here on participants interested on OGD which represent a very specific sub-set of the population, (ii) we only used online channels due to the COVID situation and (iii) the survey was conducted as a complement to the usage statistics.

2.2 Interviews (RQ2)

We conducted interviews with 10 non-expert citizens of Namur, interested to know more about OGD, to answer RQ2. These 10 citizens were recruited on

[1] https://rb.gy/7sgpqo
[2] https://data.namur.be/pages/accueil/
[3] https://data.overheid.nl/community/maatschappij/high-value/gemeenten
[4] https://rb.gy/2wjtgd

voluntary basis based on their previous answers in the survey described in Section 2.1. The reasons for low participation are the same as above, with the exception of the third. NeDVis was used to facilitate the data collection for the interviews. This tool was selected because it allows to easily take into account the user feedback compared to the existent solutions.

Datasets & predefined tasks. We selected three datasets from the open data portal of Namur for the interviews. These datasets were chosen because they are easy to understand by participants and also are among the most visited datasets according to the usage statistics file collected from the OGD manager of Namur. For each dataset, we have defined 2 tasks that the participants need to do in order to record their feedback. The predefined tasks were also well selected in order to cover different use cases of data visualization. All the datasets and predefined tasks were later integrated in NeDVis in order to facilitate the interview process and especially to not lose time during the interview. Table 1 summarizes the information about the datasets (name, link for more details and predefined tasks).

Table 1. List of datasets and predefined tasks for interviews.

Datasets	**Predefined Tasks**
COVID-19 Pandemic - Province of Namur - New contaminations by commune Link: https://rb.gy/4r0ht7	(T1) Total new cases over date (T2) Total new cases per municipality
Namur - Mobility - Parking Link : https://rb.gy/b840w5	(T3) Total places per parking type (T4) Total places per parking type and per municipality
Namur - Ordinary budget by function Link : https://rb.gy/1q79nd	(T5) Total revenues and total expenses across function (T6) Total revenues and total expenses across function over year

Data collection. In order to reduce the duration of the interview, we launched NeDVis on our computer and asked participants to evaluate the generated visualizations. For each predefined task, NeDVis generated at least 2 different visualizations. Participants were then asked to give a score between 1 (very inappropriate) to 10 (very appropriate) to each generated visualization. In addition, participants were asked to verbalize their thoughts during the study about why they gave a certain score for a specific visualization and also how they would like the system to represent the visualization to facilitate understanding. These thoughts were recorded so that nothing was missed from their feedback. Each subject spent approximately 30min to note in total 22 visualizations for all the predefined tasks.

Data analysis. After collecting user feedback, the final score of each proposed visualization type for each predefined task was calculated using the average of user ratings. The different scores were then used to find the best visualization type (visualization with the highest score) for each predefined task.

3 Results

3.1 Content expectations: usage statistics and survey results

The file collected from the OGD manager of Namur portal concerns the consultation of the portal's data from January to December 2020. This file contains 902

573 rows and 34 columns such as timestamp, user_ip_addr, dataset_id, exec_time and so on. Based on this file, we determine how many times each dataset was visited between January and December 2020. Table 2 shows the top 10 datasets consulted on the OGD portal of Namur.

Table 2. Top 10 datasets visited between January and December 2020 on the OGD portal of Namur.

No	Dataset	Dataset Category in Survey	Number of records	% records
D1	Number of confirmed COVID-19 cases by municipality	COVID	298498	33.1
D2	Number of new confirmed COVID-19 cases per municipality per day	COVID	57130	18.4
D3	Number of new hospitalizations of COVID-19 per province per day	COVID	51232	6.33
D4	List of Deceased Related to Cemetery Locations	Non Present	24260	5.68
D5	Administrative boundaries - Municipalities of the Province of Namur	Non Present	20108	2.69
D6	Polygons of 26 localities of the commune of Namur	Non Present	17776	2.23
D7	Boundaries of districts of Namur	Non Present	16345	1.97
D8	Location of Public Cemeteries	Non Present	14588	1.81
D9	Photos and geolocalized old postcards	Non Present	10497	1.62
D10	List of the deceased linked to the cemetery sites in the commune of Namur	Non Present	8756	1.16

Referring to Table 2, the order of the datasets from most to least visited, is as follows: COVID datasets, datasets on cemeteries, data on communities, localities, addresses, and buildings, mobility data and population data. Another observation is that some expected datasets, such as budget data to achieve transparency, were less visited but were among the 100 most visited datasets. Also, many datasets in the list of HVDs are not found in the list of datasets visited on the Namur portal.

Regarding the survey, a total of 43 users completed it. 63% of users had heard about open data, 53% had used an open data portal and 70% had general computer knowledge. First, we asked participants to quantify the importance of the predefined datasets (coming from the list of HVDs) using a scale from "Not important at all" to "Very Important". The importance was calculated as the median response of the 43 respondents. The survey results show that most of the datasets are important (median=3) for citizens except the datasets about street lighting, places to walk dogs, information on trees and spreading routes, which have a median less than 3 (not important). Second, we asked the following question to participants: "What data (other than those listed) would you like to see on an Open Data site?". 13 participants answered it. The list of suggested data included: nurseries libraries, road work schedule, local business statistics, position of the refugee centers and their age pyramid, collection and use of tax and information on essential shops.

Based on these findings, we suggest publishers to highlight on the portal (respectively developers to offer services based on) the high-value datasets, COVID-Related Data (or, more generally, data relevant to analyze a current crisis and/or societal debate in an objective manner), administrative boundaries and popu-

lation data, a list of buildings, mobility data and old photos from the city. On the other hand, publishers should also provide, in addition to the current data, datasets about nurseries libraries, road work schedule, local business statistics, position of the refugee centers and their age pyramid, collection and use of tax and information on essential shops. They should also have a feedback feature which can help collect user expectations in terms of the datasets to be published on the portal.

3.2 Visualization expectations: results from interviews

In total, 10 users participated to the interviews. All participants had average to low computer skills and had not previously analyzed the studied datasets. Table 3 presents the adequate visualization type for each predefined task based on the user feedback. Note that the best visualization type is determined by taking the visualization type that has the highest final score calculated using the average of user ratings.

Table 3. Best visualization type for each predefined task.

Datasets	**Tasks Predefined**	**Best Visualization Type**
COVID-19 Pandemic - Province of Namur - New contaminations by commune	(T1) Total new cases over date	Line chart
	(T2) Total new cases per municipality	Bubble map
Namur - Mobility - Parking	(T3) Total places per parking type	Bar chart with horizontal orientation & Doughnut & Pie chart
	(T4) Total places per parking type and per municipality	Grouped bar chart with horizontal orientation
Namur - Ordinary budget by function	(T5) Total revenues and total expenses across function	Grouped bar chart with vertical orientation
	(T6) Total revenues and total expenses across function over year	Multiple line charts

Referring to Table 3, we can note that the best visualization type for visualizing geographic data is the bubble map, for comparing categorical data is the bar graph, and for seeing the evolution over time is the line graph. In addition, we find that the design of the visualization types is very important for non-experts to help them understand them easily. Thus, based on user feedback on suggested visualizations, we propose that programmers and publishers take the following actions to incorporate these user expectations. First, they should have a visualization review feature that allows users to provide suggestions on how to improve visualizations. Second, they should allow users to access the low-level visual encodings such as graph orientation, axis labels, order of data in graph and color, in order to change them if necessary. Third, they should provide search functionality for each visualisation to allow only the desired data to be displayed rather than all data.

4 Conclusion and Further Research

The aim of this paper was to understand the content (RQ1) and visualization expectations (RQ2) of non-experts towards OGD. To achieve that goal, we used

a multi-method approach including an analysis of the usage statistics of the OGD portal of Namur, a complementary online survey of 43 participants to find out the needs of the end-users in terms of datasets and interviews with 10 participants to get their opinion on the correct and well-designed visualizations of datasets. Using this multi-method approach, we identify end-users' expectations for content and visualizations, and then provide useful recommendations to programmers and publishers. This study differs from existing literature in two aspects. First, to our knowledge, this study is the first attempt to use the usage statistics of portal combined with a survey to understand content expectations. Second, in previous researches, only few visualization types were used [6] and general interactivity (not based on tasks and feedback from citizens as done here) were suggested [7]. For future work, we first plan to increase the number of participants and cover more visualization types in order to have statistical significance. Second, we plan to improve NeDVis tool by integrating recommendations gathered from the interviews: (A) which type of visualization is appropriate for a specific task? (from scores), and (B) how to represent the visualization to make it easier to understand? (from verbal thoughts). The integration of (A) will be handled by recording the score and related features(e.g., visualization type, number of numerical attributes) of each rated visualization in the system, which will help improving the recommendation module of NeDVis. For (B), we will need to modify the representation of some visualization types and also allow users to make changes to low-level visual encodings when selecting an attribute to visualize.

References

1. Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R. S.: Socio-technical Impediments of Open Data. Electron. J. Electron. Gov. 10, 156–172 (2012).
2. Tammisto, Y., Lindman, J.: Definition of open data services in software business. in Lecture Notes in Business Information Processing vol. 114 LNBIP 297–303 (2012).
3. Crusoe, J., Simonofski, A., Clarinval, A., Gebka, E.: The Impact of Impediments on Open Government Data Use: Insights from Users. 13th International Conference on RCIS. 1-12 (2019).
4. Barcellos, R., Viterbo, J., Miranda, L., Bernardini, F., Maciel, C., Trevisan, D.: Transparency in practice: using visualization to enhance the interpretability of open data. 18th Annual International Conference on Digital Government Research. 139-148 (2017).
5. Grimm, P: Pretesting a questionnaire. Wiley Int. Encycl. Mark. (2010).
6. Ornig, E., Faichney, J., Stantic, B.: Empirical Evaluation of Data Visualizations by Non-Expert Users. 10, 355–371 (2017).
7. Khan, M., Shah Khan, S.: Data and Information Visualization Methods, and Interactive Mechanisms: A Survey. Int. J. Comput. Appl. 34, (2011).