Participation and Data Quality in Open Data use: Open Data Infrastructures Evaluated

Anneke Zuiderwijk, Marijn Janssen
Delft University of Technology, Delft, The Netherlands
a.m.g.zuiderwijk-vaneijk@tudelft.nl
m.f.w.h.a.janssen@tudelft.nl

Abstract: Infrastructures may improve the use of Open Government Data (OGD) by providing insight in how individuals can participate in data reuse and in the quality of open data. Yet, most OGD infrastructures do not support such activities. The objective of this paper is to evaluate the importance and usability of participation mechanisms and data quality indicators for open data infrastructures through quasi-experiments. A quasi-experiment is an experimental evaluation method in which researchers have control over the (non-random) assignment of participants to treatment and control conditions, the selection of control conditions and the organisation of the treatment, and over the measures. Moreover, quasi-experiments usually include a pre-test (i.e. a test before the treatment or control condition) and a post-test (i.e. a test after the treatment or control condition). The results of our quasi-experiments showed that the prototype of our OGD infrastructure improved OGD participation and data quality analysis by providing functionalities including discussion messages, social media sharing, linking items related to a dataset, Wiki descriptions and discussions, and data quality ratings and reviews. Participant observations showed that participants in the treatment group found it easier to conduct tasks with the prototype related to giving feedback on and discussing open data and rating and reviewing data quality than the participants in the control group. Our study suggested that participation mechanisms and quality indicators add value and improve the use of OGD. It recommends the implementation of such mechanisms and indicators in existing OGD infrastructures. To support the creation of transparency, citizen participation and innovation with OGD, our findings suggest that participation mechanisms and data quality indicators are a condition. Yet, these mechanisms and indicators are not sufficient for ensuring the generation of the OGD benefits, since there are still many factors which hinder the generation of these benefits. We discuss a number of these factors including factors related to OGD infrastructures and factors beyond OGD infrastructures.

Keywords: open data, participation, data quality, usability, experiments, infrastructure, social media

1. Introduction

Governmental organisations increasingly open their data to obtain benefits such as transparency (Bertot et al., 2010, Kassen, 2013), citizen participation (Conradi and Choenni, 2014) and innovation (Janssen, 2011). The literature indicates that the use of the Open Government Data (OGD) can be the starting point for democratic dialogues (Davies, 2010), and that data providers can use information about OGD use to make more informed future investment decisions concerning the supply of open data (Davies, 2010). Participation may take place by allowing citizens to contribute to discussions on how to better address their needs (Kassen, 2013). For instance, social media technologies allow for access to and interaction with government operations, programs and data (Bertot et al., 2012). Social media can be used to stimulate participation (Veljković et al., 2014) and to engage people in open data (Garbett et al., 2011). The quality of the data plays an essential role in the use of government portals (Dettlor et al., 2013), and a certain level of data quality is critical for OGD use (O’Hara, 2012).

However, at present there is limited support for the use of OGD, while the use and the exploitation of OGD is required to obtain the envisioned benefits (Jurisch et al., 2015). Research on OGD in particular has confirmed that a first wave of OGD infrastructures mainly provides basic functionalities for uploading and downloading data (Alexopoulos et al., 2013, Charalabidis et al., 2014). Existing OGD infrastructures often lack opportunity for data users to participate in improving published data (Alexopoulos et al., 2013). Access on itself is not enough to generate active participation (Alani et al., 2008). The literature shows that participation in OGD use is limited, for example because conversations about released data are lacking (Lee and Kwak, 2012) and because many OGD providers do not know who their external users are (Archer et al., 2013). Users may
also be concerned about the quality of open data (Martin, 2014). Kuk and Davies (2011) state that open data often suffer from poor quality, such as inconsistency in terms used in datasets and a lack of granularity. There is a wide variety in the quality of the released datasets (Auer et al., 2013, Kuk and Davies, 2011, Petychakis et al., 2014).

Citizen engagement through user participation and through the provision of high-quality data is considered to be important in open data use to accomplish benefits such as innovation and participation. However, there is hardly any research about the actual engagement effects. The objective of this paper is to evaluate the usability of participation mechanisms and data quality indicators for OGD infrastructures. This research uses quasi-experimentation as the main research instrument to test whether the inclusion of participation and quality mechanisms results in improved OGD use.

2. User engagement functionalities

Participation can be defined as “including citizens in the process of forming the state”, and citizens may propose directions for the development of state services and guidelines (Parycek and Sachs, 2010, p. 3). In other contexts it has been argued that a culture of participation through collaborative citizen and government networks may lead to participation in public agenda-setting and decision-making (Maier-Rabler and Huber, 2011). This may also apply to the OGD domain. Open data research generally suggests that participation mechanisms can be beneficial to OGD users. A diversity of participation mechanisms may affect the extent to which users can participate on OGD portals and the extent to which these portals can be used to engage in collaborative innovation (Sayogo et al., 2014). For instance, the analysis of feedback from OGD users can help in improving the procedures for newly publishing or updating datasets (Kucera and Chlapek, 2014), which may lead to continuous improvements to datasets of benefit to all future users of the dataset (Dawes and Helbig, 2010).

Furthermore, open data success strongly depends on the quality of released datasets (Behkamal et al., 2014). Information quality is determined by various indicators, such as accuracy, completeness, consistency and timeliness (Batini et al., 2009). Trust in the quality of datasets is critical for OGD reuse (O’Hara, 2012). However, the quality of data varies widely (Kuk and Davies, 2011, Petychakis et al., 2014). When the quality of OGD is low, open data users may be concerned about the quality of the data (Martin, 2014). Auer et al. (2013) emphasise the importance of generating strategies for assessing the quality of data published on the Web.

We propose that participation mechanisms and data quality indicators can assist in improving OGD use. To be able to evaluate the usability of participation mechanisms and data quality indicators, we developed a prototype of an OGD infrastructure which incorporates these mechanisms and indicators. Based on the literature and discussions with potential OGD users, various prototype functionalities have been implemented (see Table 1).
Table 1: Prototype functionalities investigated in this paper.

<table>
<thead>
<tr>
<th>OGD infrastructure element</th>
<th>Functionality</th>
<th>Functionality description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation mechanisms</td>
<td>1. Discussion messages</td>
<td>Posting messages to discuss a dataset or to discuss conclusions based on data use (e.g. users can describe how they used a dataset and what they learned from this). For each message it is visible who posted it and there was some basic information about the background of the user</td>
</tr>
<tr>
<td></td>
<td>2. Social media sharing</td>
<td>Sharing a dataset or findings from data use via social media (e.g. Twitter, Facebook, LinkedIn). OGD users and providers can inform each other about what they did with and learned from a dataset</td>
</tr>
<tr>
<td></td>
<td>3. Submission of related items</td>
<td>Submit an item related to the original dataset (e.g. a publication that was written based on the dataset, a report about the data collection method or a visualisation or application of the dataset)</td>
</tr>
<tr>
<td></td>
<td>4. Wiki descriptions and discussions</td>
<td>Using a wiki to describe and discuss OGD use in general. The wiki contains documentation and tutorials about how the prototype can be used to visualise and curate datasets</td>
</tr>
<tr>
<td>Data quality indicators</td>
<td>5. Data quality ratings</td>
<td>Viewing how other users assessed the dataset (i.e. viewing ratings of various quality indicators of the dataset) and rating the quality of the dataset</td>
</tr>
<tr>
<td></td>
<td>6. Data quality reviews</td>
<td>Discussing the quality of the dataset by leaving a message or review in the data quality rating system, also visible to other OGD users</td>
</tr>
</tbody>
</table>

3. Evaluation methodology: Quasi-experiments

The evaluation aimed at examining the usability of participation mechanisms and data quality indicators. The following sections describe the quasi-experimental approach of this research, as well as the organisation of the quasi-experiments.

3.1 Quasi-experimental approach

Generally, experiments can be carried out to manipulate variables and to examine their effects on other variables (Campbell and Stanley, 1969). Experiments can be either true experiments or quasi-experiments. Quasi-experiments incorporate a treatment and a control group, a pre-test and a post-test, and a model that reveals the treatment and the control group effects over time, given no treatment effects (Kenny, 1975). They allow researchers to hold control over the selection and scheduling of measures, the non-random assignment of participants, the type of control group to compare the treatment group with, and how the treatment is organised (Shadish et al., 2002). Our experiments did not allow for randomly assigning participants to treatment and control conditions and thus we cannot refer to them as true experiments (Campbell and Stanley, 1969). We therefore performed quasi-experiments.

Figure 1 provides an overview of the variables involved in our quasi-experiments. The benefits on the very left side of the figure are outside the scope of our evaluations, because they are too high level to evaluate through observations. For example, transparency is a construct that needs to be measured through multiple variables. We decided to focus on more tangible variables, namely on the infrastructure elements. Yet, we will reflect on the high level benefits of transparency, participation and innovation in our discussion (section 5).
3.2 Organisation of the quasi-experiments

In 2014 three quasi-experiments were conducted with 19 third year Bachelors students, 72 first year Masters students, and 36 professional open data users. Each quasi-experiment took 95-100 minutes. Participants of the first and second quasi-experiment were randomly divided in a treatment group (i.e. a group that used the designed prototype) and a control group (i.e. a group that used a control OGD infrastructure) (see Figure 2). The participants from the third quasi-experiment were not split into two groups because the size of this group did not allow for this. In the first and second quasi-experiment the participants were provided with two different infrastructures to test the effects of the introduction of participation mechanisms, and data quality indicators. The other conditions for the treatment groups and the control group remained as equal as possible. The control OGD infrastructure did not provide participation mechanisms. The control OGD infrastructure did provide rates about quality aspects of datasets (e.g. about the completeness of the data and the format that is was provided) for a number of datasets. There was no possibility to write a free-text review about the quality of the data or for which purposes it could be used. In both the control and the treatment condition in all quasi-experiments most participants were male (71.7% males in total). The average age of the 120 participants who provided age information was 27.9 years, ranging from 20-65. Most quasi-experiment participants (62.2%) were Dutch.

All three quasi-experiments comprised a practical session in which the participants worked with the OGD infrastructure by conducting the same scenario tasks. The quasi-experiment participants received instructions in advance. A pre-test (a survey) was carried out to measure various background characteristics of the participants, as well as their experience with OGD use. Subsequently, the participants completed scenario tasks to evaluate the infrastructure functionalities, as well as a second survey about the difficulty of these tasks. While the participants completed the scenarios tasks, time measures and observations were used to obtain additional information. An observation protocol was developed, provided to the observers, and explained to them in a training session. To structure the observations and make the observers focus on the infrastructure functionalities, observers were provided with a semi-structured observer survey. Then a post-test (the third survey) was conducted to measure whether the OGD infrastructure had influenced to which extent the scenario tasks could be completed. This paper discusses the results from the observations.
4. Measuring OGD use: findings from the observations

In this section we describe the results from the participant observations, which were used to investigate to which extent participation mechanisms and data quality indicators influenced the ease of OGD use. Observers were asked to assess the ease of conducting the scenarios by the participants of the quasi-experiments. Table 2 gives an overview of the assessment of the quasi-experiments by the 18 observers. In general it was found that the observers believed that the control group participants found it more difficult to conduct the scenarios than the treatment group participants. We elaborate on the findings from the observations in the following sections.

Table 2: Assessment of the ease/difficulty of the scenarios according to the observers (the numbers represent the number of observers).

<table>
<thead>
<tr>
<th>Scenario: OGD participation</th>
<th>Very difficult or difficult</th>
<th>Not difficult, nor easy</th>
<th>Easy or very easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group (students)</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Treatment group (students)</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Treatment group (professionals)</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario: OGD quality analysis</th>
<th>Very difficult or difficult</th>
<th>Not difficult, nor easy</th>
<th>Easy or very easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group (students)</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Treatment group (students)</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Treatment group (professionals)</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

4.1 Observations of OGD participation scenario

The OGD participation scenario encompassed giving feedback on and discussing data use. Participants of the quasi-experiments could post messages, discuss data use on social media and a Wiki and submit an item related to the original dataset (e.g. an improved or extended dataset). The observers of the control group observed that it was difficult or very difficult for the participants to use the control infrastructure for these purposes. The observers of the control group wrote that “participants found it quite difficult [...] and they gave up too quickly”. The observers of the treatment group of students observed that it was easier for the students to conduct the participation scenario. Most of these participants could find the functionalities that they needed to participate in OGD discussions, although it still took them some time: “they are trained to form opinions but taking the task seriously they took time to formulate their opinion well”. One observer mentioned the difficulty for the students in the treatment group to find the Wiki: “the students can easily find the discussion under the selected dataset, but it is difficult for them to find a forum or Wiki related to the dataset”. The observers of the treatment group of professionals all stated that the participants found it easy or very easy to conduct the participation scenario (“it’s really easy for them to look at other people’s comments under the specific dataset”). The professional OGD users may be more trained to participate in OGD use and to formulate their opinion regarding various aspects of the dataset.

4.2 Observations of OGD quality analysis scenario

The subsequent scenario concerned the analysis of OGD quality through ratings and reviews. Participants could view how other users assessed the dataset and they could discuss the quality of the dataset by leaving a message or review in the data quality rating system. The observers indicated that it was relatively difficult for the control group participants to conduct this scenario, as illustrated by one control group observer who said that “all of them could not finish the tasks”. Another control group observer stated that “only some of them saw the rating indexes. No one has been seen to leave a message”. It was observed that it was difficult to use the control OGD infrastructure to rate and review the quality of open datasets. On the other hand, the observations showed that the student and professionals of the treatment groups found this scenario less difficult to complete. The observers of the treatment group wrote that this was an “easy task for all the participants” and that “even the slowest participant did it in a few minutes”. Another observer mentioned that “students find it easy to rate the dataset”. It was observed by all six observers of the treatment group of professionals that it was easy or very easy to use the developed OGD for data analysis purposes. While five other observers also mentioned that students in the treatment group found it easy or very easy to use the developed OGD for data analysis purposes, one observer indicated that the observed students found this
difficult or very difficult. It is not clear why this observer disagreed with the other observers of the student and professional treatment group.

4.3 Observations of intermediate variables

Various intermediate variables were investigated through the observations. Intermediate variables can be defined as variables which might influence the effect of the independent on the dependent variables (Pearl, 2001), yet it is not clear whether this influence actually exists and what its nature is. The observers were asked to which extent they believed that various other intermediate variables besides the elements of the infrastructure may have influenced the difficulty or ease of performing the scenario tasks in the quasi-experiments. We first discuss the intermediate variables observed by the observers of the control group, followed by those observed in the student treatment group. Thereafter, the observation of intermediate variables in the professional treatment group is discussed. Finally, we discuss intermediate variables for all three groups.

First, out of the five observers of the control group, four wrote that the performance of the control group had negatively been influenced by the user interface. It was mentioned that the user interface of the control infrastructure was not user friendly and could be improved. Furthermore, two observers wrote that some of the observed participants might have been influenced slightly by the observers, since it was visible to the participants that they were observed. Four observers wrote that some participants may to some extent have influenced other participants, and one wrote that previous experience had influenced their performance (“their previous experience was important”). Moreover, one observer wrote that the setting may have influenced the performance of the participants because it was difficult for some participants to see the screen with instructions, although these participants could still hear the instructions as presented by the facilitator of the quasi-experiments. Finally, one observer wrote that the temporal unavailability of the server of the control group had frustrated some participants. No influence was observed from the facilitator, the nationality of the participants or the organisation of the quasi-experiment.

Second, intermediate variables were investigated for the student treatment group. All six observers of the student treatment group wrote that the difficulty or ease of performing the scenarios had been influenced by the user interface (“I believe that the ease and the simplicity of the user interface played a major role”). This finding indicates that the usability of the OGD infrastructure may not only be the result of the infrastructure elements, but that the user interface also plays an important role. Since the observers of the control group also found that the user interface could negatively influence the usability of the infrastructure, it is important for OGD infrastructures to devote sufficient attention to the development of the user interface. Two of these six observers of the student treatment group wrote that some participants might have been influenced slightly by the observers, since the participants could see that they were observed. Two observers stated that some participants may to some extent have influenced other participants, as some asked each other questions. Two observers wrote that the nationality of two participants may have made it slightly more difficult to conduct the tasks for non-Dutch participants, which may have been caused by the fact that the participants had to reuse a dataset (in English) that concerned Dutch elections. One observer stated that some noise of participants who had finished the quasi-experiment may have influenced the performance of other participants who were still working on the scenarios. It was observed that the organisation of the quasi-experiment was clear to almost all participants. One observer said that “one person didn’t understand some instructions”, while other observers stated that there was no influence from the organisation of the quasi-experiment at all (“everything was well-prepared, participants were not distracted by unclear aspects”). No influence was observed from the facilitator of the quasi-experiment, or the experience of the participants in the student treatment group (“their e-skills are so high, they immediately got used to the interface”).

The third group that was observed in the quasi-experiments included the treatment group of professional OGD users. Out of the seven observers of the professionals treatment group, five wrote that the difficulty or ease of performing the scenarios had been influenced by the user interface. This observation confirms the observations of the participants of the control and the student treatment groups, and it shows the importance of paying considerable attention to the user interface to improve OGD use. Moreover, three observers wrote that the participants’ experience had influenced their performance (“more experienced participants performed better and faster”). This finding suggests that OGD use can be improved by training potential users, so that they obtain experience with the use of OGD infrastructures.

Two observers stated that there had been a slight influence from the facilitator at the end of the quasi-experiment with professionals after the facilitator had announced that time for conducting the scenarios had almost passed. This had resulted in rushing through the final scenario by some participants. Some
observers said that older participants worked more slowly than younger participants. There were no significant differences observed between males and females who participated in the quasi-experiments. In addition, no influence was observed from the nationality of the participants or the setting of the experiment. Almost no influence was observed from the organisation of the quasi-experiments (“Well-structured organization. No guidance questions from the participants”). No influence was observed from the observers in the treatment group of professionals, and only one participant was found to be influenced by other participants. In all quasi-experiments the computers were separated by partitions, so that all the participants had their own work place and it was difficult for them to see what other participants did on their computers. This stimulated working individually.

Finally, the infrastructure also contained functionalities that were not described in this paper. For instance, it was also possible to visualise datasets in charts and tables, and structured overviews of the metadata of each dataset were available to the infrastructure user. These other functionalities may also have influenced the ease or difficulty of conducting the scenarios.

5. Discussion and conclusions

The objective of this paper was to evaluate the usability of participation mechanisms and data quality indicators for open data infrastructures. Participant observations in three quasi-experiments showed that participants in the treatment group found it easier to conduct tasks with the infrastructure prototype related to giving feedback on and discussing open data and rating and reviewing data quality than the participants in the control group. The study suggested that the participation mechanisms and quality indicators add value and improve the use of OGD. We recommend the implementation of such mechanisms and indicators in existing OGD infrastructures.

This paper started with the statement that OGD are often released for the reason to stimulate benefits including transparency, citizen participation and innovation. The findings of this study suggested that participation mechanisms and data quality indicators are essential to support transparency, citizen participation and innovation with OGD. Without participation mechanisms it is not possible for OGD users and providers to discuss what can be learned from the use of certain datasets, and OGD users cannot give feedback on datasets or data reuse. Without indicators of OGD quality, potential data users do not know for which purposes a dataset can be used, and they may be reluctant to actually use the dataset.

Nevertheless, participation mechanisms and data quality indicators are not sufficient for ensuring the generation of transparency, citizen participation and innovation. Various other factors hinder the generation of these benefits. Some of these hindering factors concern other infrastructure functionalities. For example, OGD infrastructures also need to provide insight in the context in which datasets have been created and sufficient contextual metadata needs to be provided to OGD users. While the literature postulates that it is essential for the correct interpretation and use of open data to offer sufficient metadata simultaneously to data (Braunschweig et al., 2012), metadata provision for open data is often cumbersome (Martin, 2014, Dawes and Helbig, 2010). Another infrastructure barrier concerns the lack of tools to search, analyse, visualise, mine and otherwise obtain understanding of open datasets. OGD users can be supported through integrated adequate and context-relevant knowledge processing tools and resources (Charalabidis et al., 2011). However, most traditional open data infrastructures provide only basic data download and upload functionalities (Alexopoulos et al., 2013, Charalabidis et al., 2014). There is a lack of tools to generate information that can easily be understood by the general public (Novais et al., 2013).

Infrastructures have large potential to support OGD access and public debate (Charalabidis et al., 2011), and they may be improved to remove the above-mentioned barriers. At the same time there are also barriers which do not relate to OGD infrastructures but to other aspects. Such factors, for example, relate to the lack of availability and access to the desired datasets. The release of OGD is challenging, and OGD infrastructures cannot fully support the internal decision-making processes that need to take place within governmental organisations before their data are disclosed. While some researchers have proposed guidelines and decision-support for releasing governmental data (e.g., Zuiderwijk et al., 2014), data publishers still need to handle obstacles related to threats of privacy violations by releasing data, of being legally liable when disclosed data are misused (Kalidien et al., 2010, Kulk and van Loenen, 2012), and of potential misinterpretation and misuse of their datasets (Conradie and Choenni, 2014, Kucera and Chlapek, 2014). Datasets that are desired by potential OGD users may not be available or accessible.

OGD infrastructures can be improved further to enhance their usability and to stimulate the high-level benefits such as transparency, innovation and participation. In addition, other factors which hinder the realisation of OGD benefits need to be addressed. Transparency, participation and innovation require an
integrated approach towards removing OGD barriers, and factors including those related to infrastructures (including participation mechanisms, data quality indicators and metadata), decision-making processes, availability and access to data, usability and understandability need to be addressed.

References


Pearl, J. (2001) "Direct and Indirect Effects", Paper read at 17th Conference on Uncertainty in Artificial Intelligence, Seattle, U.S.A.


