Open Data in Regions from the Users' Perspective: an Analytical Study

Miloš Ulman[1], Edita Šilerová[2], Jan Masner, Michal Stočes[2], Pavel Šimek[2], Petr Benda[2]

[1]Department of Information Technology, Faculty of Economic and Management, Czech University of Life Sciences in Prague, Kamycka 129, 165 21 Prague Suchdol, Czech Republic, e-mail: ulman@pef.czu.cz
[2]Department of Information Technology, Faculty of Economic and Management, Czech University of Life Sciences in Prague, Kamycka 129, 165 21 Prague Suchdol, Czech Republic

Abstract. The paper is focused on open data in regions from the users' perspective and presents the perceived usability of various public data sets in the Czech Republic. There were 265 data sources assessed by respondents according the format, suggestions of further use, number of views, and usability for citizens, businesses, officers and other users. Only 14 datasets from national public agencies were already provided as open data, but none at the local level. The most frequent formats of data were DOC, HTML, PDF and XLS. The respondents came with 36 different suggestions for the further use of data sets. Currently, the citizens cannot see the difference in usability of open and non-open data unless there are particular applications available. Nationwide data sets were assessed in usability on average better (1.57) than local data sets (1.93). The usability of data was evaluated similarly across observed regions beside those provided by national institutions.

Keywords: Open data, open government data, user's perspective, region, local government, usability.

1 Introduction

Open data is the term that echoes from local and world media with growing frequency. Open data are complete, easily accessible, machine readable, using open standards (e.g. CSV or XML) and published with an open license (Auer et al, 2007). The term of open data firstly gained popularity in academia where it denotes effort to publish academic data under free access in special digital depositories (Murray-Rust, 2008). Now the idea of open data is mainly perceived with political meaning, especially, due to the launch of government open data portals data.gov and data.gov.uk in the U.S. and the U.K. (Kassen, 2013) and with the start of initiatives such as Open Government Partnership. Making public sector information freely accessible in open formats is also referred as *open government data (OGD)* (Kalampokis et al, 2011; Shadbolt et al, 2012).

Copyright © 2015 for this paper by its authors. Copying permitted for private and academic purposes.
Proceedings of the 7th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2015), Kavala, Greece, 17-20 September, 2015.

Benefits and advantages of using open data provided by government and public agencies (or open government data) are promising for the public administration and for private sector (World Bank, 2014). But most public organizations have no or limited interaction with data users and are often very selective in communication (Susha et al, 2015). Some authors propose that open data rather should be a way for government to interact with citizens (Sieber & Johnson, 2015).

The Czech Republic has committed to implement open government data by joining the Open Government Partnership in 2011. Since then a number of OGD activities have been started at different levels of the public sector as well as in academia and other domains. Although there are some challenges that the Czech Republic needs to face – like the missing official OGD catalogue, some public sector bodies have already started to publish open government data (Chlapek et al, 2014). There are several pioneering organizations that started as first and have showed already some tangible results. Among these organizations belong national state agencies (namely the Ministry of Finance, Czech Trade Inspection or Czech Telecommunication Office), regional administrations (e.g. Vysocina region) and municipalities (e.g. Prague, Decin, Opava). The national open data catalogue of the Czech Republic has been recently launched within the main government portal portal.gov.cz. There are concentrated all datasets from national public institutions that have opened their data so far. The main resource is the register of public contracts that provides scanned copies of all public tender contracts with metadata in XML for automated processing.

Enormous data volumes are generated on daily basis by municipalities and regional authorities. The demand for public data in open format and the number of relevant applications are expected to be steadily growing as an effect of the increase in the level of digital skills and demand for information for example among inhabitants in rural areas (Vaněk et al, 2011) and small farmers (Kubata et al, 2014). The extend of open data sources could have a positive impact on areas such as tourism (Šilerová, 2013). Some authors argue that the potential of the open data concept can be realized at the local level (Kassen, 2013). Contributions and particular impacts of open data on rural life should be examined and researched. The Department of Information Technologies FEM CULS Prague has started cooperation with Czech Ministry of Agriculture on the research of open data in agriculture and rural regions.

There are two aims of the paper: firstly, to estimate the potential of open data sources at the national and regional level, and secondly, to evaluate user's view of open data usability in the Czech Republic.

2 Materials and Methods

Particular research questions that were examined in the survey were stated such as:

1. The number of potential data sets identified for opening and the number of suggestions for the further use of open data was rather higher at local or regional administrations than at the national level.

2. There are differences in estimated usability related to the type of the municipality and the user group.
3. There is a significant difference in usability of open data and data that are not open.
4. There are differences in estimated usability related to the size of the municipality population.

Based on research questions further statistical processing was conducted. The methodological approach used in the paper is both of qualitative and quantitative nature and includes literature review, user interviews and descriptive analysis. The data were processed with association tables, descriptive statistics and tests to compare differences in values.

3 Results and Discussion

Data were collected through questionnaire form administered to citizens living in various regions. In total, there were 265 data sets assessed by respondents in terms of format, suggestions of further use, number of views, and usability for citizens, businesses, officers and other users. The perceived usability of data sets was ranked on the scale from 1 (the best) to 5 (the worst). Basic data are summarized in Table 1.

Table 1. Data sets overview from the survey.

Location	Data sets	Open data	Suggestions
Local	216	0	26
National	49	14	10
Total	265	14	36

Source: self-authored, 2015.

The collected data represent 52 municipalities from 9 regions (including the capital Prague), 8 public institutions (7 schools and 1 school dormitory) and 11 state bodies in the Czech Republic. The most frequently evaluated types of data sets were: obligatory information published on public office website (also called as “e-desk”) (33 %), budget information (33 %), annual report (13 %), newsletter (10 %) and decision of representatives (10 %). In the survey focused on the openness to information disclosure among between 395 municipalities (with up to 2,000 inhabitants) in one region of the Czech Republic in 2009 (Bachmann, 2012), it was found that 71 % of municipalities published electronically minutes from the municipal council meetings, and 28 % electronic periodicals (newsletter). In the survey between 400 municipalities across the whole Czech Republic in 2012, there were 27 % of municipalities publishing resolution from the municipal council meetings, 17 % resolutions and minutes, and 40 % of municipalities publishing newsletter. The procurement information was published only in 30 % of cases. However there is a significant correlation between the size of municipality population and its information openness (Bachmann & Zubr, 2014).

There were 16 various formats of data sets identified such as DOC, HTML, XLS, PDF, XML, CSV and other. Additional 6 various combinations of certain formats were used and some datasets were already provided in the form of a web application. Respondents came with particular ideas of further use of data sets in open format in 36 cases (13.6 %), while 14 (5.3 %) data sets were already provided according to open data principles. The remarkable finding is that all open data sets were provided by state administration bodies and none by local authorities. However more suggestions came for local data sets (26) than for national (10). The most suggestions for the further use of data were declared with files in PDF (9 suggestions), HTML (8), XLS (5), HTML combined with PDF (4) and CSV (4). The first research assumption was confirmed since the availability and the number of inputs for data sets use is prevailing at regional level.

The second question was explored by conducting a statistical analysis. Regarding the nature of gathered data differences among mean values were examined by the analysis of variance and several non-parametric tests. We found that data usability evaluation does not have normal distribution and samples are of different size. ANOVA test assumptions such as normality and symmetry of sample distribution were not fully satisfied, so non-parametric tests were also employed, namely Kruskal-Wallis, Leven's test (means) and Welch's test (Lantz, 2013; Zimmerman, 2011). The statistical hypotheses that were tested are presented in Table 2.

Table 2. Statistical hypotheses related to data sets usability evaluation

Hypotheses	
H1	There is no significant difference between the type of format of the data set and its usability (from the citizen's perspective).
H2	There is no significant difference between the local and national providers and the estimated usability (from the citizen's perspective).
H3	There is no significant difference between regions and the estimated usability (from the citizen's perspective).
H4	There is no significant difference between the evaluation of the data set usability and the user group.
H5	There is no significant difference in usability between data provided as open data and non-open data.
H6	There is no significant difference in usability between municipalities according to size of population.

Source: self-authored, 2015.

Basic descriptive statistics and results of hypotheses testing are presented in following tables (Table 3 – 12).

Firstly, we focused on differences between data format and perceived usability of data. Based on results in Table 4, we can conclude that there are no significant differences in the means and the type of format has no influence on the perceived usability of data. It is needed to add that only three data formats were included because the other had less than 10 evaluations that might affect reliability of results. Selected formats (DOC, HTML, PDF, XLS) are also typical for publishing data on web, however technically, PDF is considered to be one-star data according to the five-star data concept (Berners-Lee, 2010).

Table 3. Formats and usability from the citizen's perspective - descriptive statistics.

Groups	Count	Sum	Mean	Variance	SS	Std Err
DOC	17	35	2.0588	0.6838	10.9412	0.2284
HTML	81	150	1.8519	0.9528	76.2222	0.1046
PDF	77	144	1.8701	0.7724	58.7013	0.1084
XLS	29	58	2	1.125	31.5	0.1749

Source: self-authored, 2015.

The interesting finding brings the frequency analysis of formats usability evaluation. Data sources were assessed with 1 or 2 in usability in particular formats: DOC (64.7 %), HTML (75.9 %), PDF (80.5 %) and XLS (75.9 %). So having data in one of these formats is likely to be perceived positively.

Table 4. Formats and usability from the citizen's perspective – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's
p-value	0.7777	0.6798	0.7671	0.5844
significant	no	no	no	no

Source: self-authored, 2015.

Many examples of successful open data applications are based on data sets provided by nationwide authorities such as police or national health agency, but we suppose that there are myriads of opportunities to exploit data from local or regional resources. The respondents spotted 216 various local data sources (see Table 5). However, none of local data sets was served in an open data format. There were some significant differences in average values (and medians according to the Welch's test) of usability between national and local data sets (see Table 6). Thus, the hypothesis no. 2 (H2) has to be rejected.

Table 5. Usability of local vs. national open data from the citizen's perspective - descriptive statistics.

Groups	Count	Sum	Mean	Variance	SS	Std Err
Local	216	418	1.9352	0.9260	199.0926	0.0653
National	49	77	1.5714	0.9063	43.5000	0.1372

Source: self-authored, 2015.

The evaluation of the usability was positive (1 or 2) at local resources in 75.5 % of cases and at national resources in 89.8 % of cases.

Table 6. Usability of local vs. national open data from the citizen's perspective – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's
p-value	0.0174	0.0049	0.0185	0.8265
significant	yes	yes	yes	no

Source: self-authored, 2015.

Collected answers were categorized according the municipalities and regions where data were originated or related. In Table 7, there are four regions that provided at least 10 scores: whole Czech state administration (CZ), Olomoucký region (M), Plzeň region (P) and Ústí nad Labem region (U). The differences between regions were not statistically significant with 95 % probability (see Table 8). The other regions that were presented in responses, but not included in computation, were: South Bohemian region, Pardubický region, Vysočina region, Karlovarský region and Moravia-Silesian region.

Table 7. Regions and usability - descriptive statistics.

Groups	Count	Sum	Mean	Variance	SS	Std Err
CZ	49	77	1.5714	0.9063	43.5	0.1331
M	38	76	2	1.13514	42	0.1512
P	139	255	1.83453	0.80576	111.1942	0.0790
U	11	20	1.81818	0.56364	5.6364	0.2810

Source: self-authored, 2015.

The frequency of positive evaluation (1 or 2) of data sources usability across regions was distributed such as: national (89.8 %), Moravian-Silesian region (76.3 %), Plzeň region (78.4 %) and Ústí nad Labem region (81.8 %).

Table 8. Regions and usability – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's
p-value	0.1836	0.0972	0.2533	0.9122
significant	no	no	no	no

Source: self-authored, 2015.

The usability from the perspective of citizen, business and public administration (PA) staff was examined and summarized in tables 9 and 10. However, the results were not unanimous and we cannot simply tell whether or not there are any differences in usability evaluation. From the descriptive statistics (see Table 9), we can see that the variance and mean values from business users remarkably differ. ANOVA and Levene's tests that examine differences in means signalize that there are no significant differences, however, Kruskal-Wallis and Welch's test indicate the opposite (see Table 10). Actually, the responses were given by citizens rather than by business people or public officers. This fact might have significant impact on data.

To answer the second research question supposing difference in data usability among regions and user groups, we can conclude that the usability was evaluated similarly across observed regions and between user groups. On the other hand, the data sets provided by national institutions were ranked with better score than regional sources (see Table 5).

Table 9. Usability of open data according user groups - descriptive statistics.

Groups	Count	Sum	Mean	Variance	SS	Std Err
Citizen	265	495	1.8679	0.9389	247.8774	0.0640
Business	60	135	2.25	1.8178	107.25	0.1344
PA staff	58	109	1.8793	0.9852	56.1552	0.1367

Source: self-authored, 2015.

The evaluation of usability according user groups ranked 1 or 2 in 78.1 % (citizen), 68.3 % (business) and 79.3 % (PA staff) of cases.

Table 10. Usability of open data according user groups – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's
p-value	0.0351	0.2814	0.1204	0.0021
significant	yes	no	no	yes

Source: self-authored, 2015.

The third research question was to reveal the users' point of view on usability of data provided in open format and data that are not open yet. The descriptive summary is presented in Table 11. There were only 11 open data sets evaluated by users against 254 non-open data sets. Some significant discrepancy in mean values was confirmed only by Levene's p-value, but other test results including ANOVA prevented from rejecting the hypothesis about equality of mean values (see Table 12).

Table 11. Usability of open data and non-open data from the citizen's perspective - descriptive statistics.

Groups	Count	Sum	Mean	Variance	SS	Std Err
Open data	11	25	2.2727	2.6182	26.1818	0.2916
Non-open data	254	470	1.8504	0.8688	219.8150	0.0607

Source: self-authored, 2015.

There was also no remarkable difference in the frequency of a positive evaluation of data usability between open data (63.6 %) and non-open data (78.7 %).

Table 12. Usability of open data and non-open data from the citizen's perspective – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's
p-value	0.1574	0.7800	0.4102	9.1636E-05
significant	no	no	no	yes

Source: self-authored, 2015.

We can conclude that citizens can hardly see any difference between open and non-open data in terms of the usability. Unless there are particular applications based on open data much interest cannot be expected. The value of open data materializes only upon its use (Susha et al, 2015).

Table 13. Usability and size of municipality population – differences testing, α=0.05.

	ANOVA	Kruskal-Wallis	Welch's	Levene's	significant
Citizen	0.0292	0.0876	0.0636	0.9554	no
Business	0.1766	0.5432	0.0139	0.0002	no
PA staff	0.3064	0.4517	0.1777	0.0028	no

Source: self-authored, 2015.

The fourth research question and resulting final hypothesis were to examine differences among municipalities of different size. In the Table 13, we can see comparison of testing differences in usability between municipalities (without schools and national institutions) according to size of population. There were four different categories of the population size: under 2,000 inhabitants, between 2,000 and 4,999, between 5,000 and 9,999, and over 10,000. The categories follow guidelines of the Czech Statistical Office (CZSO, 2011). Because there was no sample where all test criteria would be significant, we could not reject the null hypothesis and we should state that there are no significant differences between mean values of the usability evaluation. The benefit from using data was not affected by the size of municipality population from the perspective of respondents. However, the respondents did not need to be residents in the respective municipalities. The number of population had some significant effect in the evaluation of municipality openness to information disclosure (Bachmann, 2012; Bachmann & Zubr, 2014).

4 Conclusion

Based on the survey, it can be concluded that data sets provided by national state agencies and authorities are more frequently in line with open data principles, while at the regional level there are still large resources that could be unleashed under open data concept. Citizens and businesses come in contact rather with local than national authorities and they could take better advantage of novel applications and services based on open data if they exist.

The main highlights of the survey findings are summarized under:

- In the sample of 265 data sets, the most frequently published data were: obligatory information published on public office website (also called as "e-desk") (33 %), budget information (33 %), annual report (13 %), newsletter (10 %) and decision of representatives (10 %);
- Out of 265 data sets, there were 216 data sets provided by local municipalities or institutions (81.5 %) and 49 by national bodies (18.5 %);
- Nationwide data sets were assessed in usability on average better (1.57) than local data sets (1.93);

- HTML (30.5 %), PDF (29.1 %), XLS (10.9 %) and DOC (6.4 %) are the most used data formats on municipality and public authority web sites. Their usability was assessed as positive (score 1 or 2) in at least 64.7 % cases;
- The usability of data was evaluated similarly across observed regions beside those provided by nationwide institutions;
- The usability of data was assessed from three perspectives: citizen (265 evaluations), business (60) and public administration staff (58);
- The average usability was the best from citizen's point of view (1.86) and the worst from the business point of view (2.25), but most of respondents were rather citizens than business people, which might have impacted results;
- Currently, the citizens cannot see the difference in usability of open and non-open data unless there are particular applications available;
- The benefit from using data was not affected by the size of municipality population from the perspective of respondents, but those respondents did not need to be the actual residents in those municipalities.

The presented results are limited due to the scope of the survey. More details about opinions of users coming from regions and local communities should be investigated. Also the willingness of people to use tools based on open data and to interact online with their local representatives should be examined thoroughly (Office, 2011; Susha et al, 2015).

Acknowledgments. The results and knowledge included herein have been obtained owing to support from the Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 20151039 "Využití otevřených dat v agrárním sektoru" (The use of open data in agricultural sector).

References

1. Auer, S., Lehmann, J., Bizer, C., Kobilarov, G., Cyganiak, R. & Ives, Z. (2007) DBpedia: A nucleus for a Web of open data. Translated from English by, 4825 LNCS.
2. Bachmann, P. (2012) Openness to information disclosure: the case of Czech rural municipalities. AGRIC. ECON. - CZECH, 58(12), 580-589.
3. Bachmann, P. & Zubr, V. (2014) What affects the information provided on the web? Case of Czech rural municipalities. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(6), 1221-1231.
4. Chlapek, D., Kučera, J., Nečaský, M. & Kubáň, M. (2014) Open data and PSI in the Czech Republic. European Public Sector Information Platform.
5. CZSO (2011) Preliminary results of the Census 2011 Population, 2011. Available online: https://http://www.czso.cz/csu/sldb/population [Accessed.

6. Kalampokis, E., Tarabanis, K. & Tambouris, E. (2011) A classification scheme for open government data: Towards linking decentralised data. International Journal of Web Engineering and Technology, 6(3), 266-285.
7. Kassen, M. (2013) A promising phenomenon of open data: A case study of the Chicago open data project. Government Information Quarterly, 30(4), 508-513.
8. Kubata, K., Tyrychtr, J., Ulman, M. & Vostrovskỳ, V. (2014) Business informatics and its role in agriculture in the Czech Republic. Agris On-line Papers in Economics and Informatics, 6(2), 59-66.
9. Lantz, B. (2013) The impact of sample non-normality on ANOVA and alternative methods. British Journal of Mathematical & Statistical Psychology, 66(2), 224-244.
10. Murray-Rust, P. (2008) Open data in science. Serials Review, 34(1), 52-64.
11. Office, C. S. (2011) Preliminary results of the Census 2011 Population. Available online: https://http://www.czso.cz/csu/sldb/population [Accessed.
12. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W. & Schraefel, M. C. (2012) Linked open government data: Lessons from data.gov.uk. IEEE Intelligent Systems, 27(3), 16-24.
13. Sieber, R. E. & Johnson, P. A. (2015) Civic open data at a crossroads: Dominant models and current challenges. Government Information Quarterly.
14. Susha, I., Grönlund, Å. & Janssen, M. (2015) Organizational measures to stimulate user engagement with open data. Transforming Government: People, Process and Policy, 9(2), 181-206.
15. Šilerová, E., Maneva, S., Hřebejková, J. (2013) The Importance of Congress Tourism for Regional Development. Agris on-line Papers in Economics and Informatics, V(3), 79-86.
16. Vaněk, J., Jarolímek, J. & Vogeltanzová, T. (2011) Information and Communication Technologies for Regional Development in the Czech Republic – Broadband Connectivity in Rural Areas. AGRIS on-line Papers in Economics and Informatics, III(3), 67-76.
17. Zimmerman, D. W. (2011) A simple and effective decision rule for choosing a significance test to protect against non-normality. British Journal of Mathematical & Statistical Psychology, 64(3), 388-409.