IMPACT ASSESMENT OF OPEN GOVERNMENT DATA

ALJAŽ FERENCEK

University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia; e-mail: aljaz.ferencek1@student.um.si

Abstract Public sector organizations produce and process increasing amounts of data and the number of research and initiatives on open data is also increasing. Defining the true value of OGD is challenging without knowing how it impacts society and its economy. While the analysis of the economic benefits of open data is one way to describe the effect of government openness, the impact of open data is measured also in social and political context. Feedback mechanisms that are currently used are mostly surveys, while the number of OGD use cases is increasing. This paper proposes a preliminary model for research on assessing impact areas of OGD in an automated manner by using text mining techniques on existing use cases.

Keywords:
open government data, impact assessment, open data, text mining, impact areas

DOI https://doi.org/10.18690/978-961-286-485-9.56
ISBN 978-961-286-485-9

1 Introduction

Open Government Data (OGD) is one of the sources of open data published by the public sector in form of databases, with the aim of promoting transparency, accountability, and creating added value (Open Government Data, n.d.). Public sector organizations produce and process increasing amounts of data, and the number of research and initiatives on open data is also increasing (Attard et al., 2016; Attard et al., 2015; Safarov et al., 2017; Ubaldi, 2013; Yan & Weber, 2018). By making their databases accessible, public institutions are becoming more attractive for political participation of citizens (Ruijer & Martinius, 2017), the creation of companies and innovative services focused on citizens is encouraged (Pereira et al., 2016), and the long-term goal is to ensure overall transparency of government information (Jaeger & Bertot, 2010). As Ubaldi (2013) points out, one of the elements for creating added value of OGD is high-impact data for the public. Currently, there are many open data policies at different levels of government, while very few systematic and structured studies have been conducted on their actual impact (Roa et al., 2019; Ruijer & Martinius, 2017; Zuiderwijk & Janssen, 2014). Many other authors (Afful-Dadzie E. & Afful-Dadzie A., 2017; Crusoe et al., 2019; Safarov et al., 2017; Wilson & Cong, 2020) also recognize the problem of measuring the impact of open data, and the reasons for that are mostly in the availability and low quality of data, costs and legal barriers and in users. Although not much attention has been paid to the impact of open data in the past, many useful solutions have emerged in recent years. As stated by Kalampokis et al. (2013), companies use open data and methods of business intelligence which help them survive in the global economy, open data helps designers to make better policies and academics to analyze data for hypotheses testing, understanding of patterns and predictions. For this reason, we recognized an opportunity to identify areas of open government data impacts from the actual use cases of open government data through text mining, and thus enable governments to create targeted and better-quality data sets.

The use of text mining is quite common in the literature, but not for the purpose of identifying impact areas of OGD. Common applications of text mining are topic modeling using Latent Dirichlet Allocation (LDA) which can be used to identify areas of OGD (Afful-Dadzie, E. and Afful-Dadzie A., 2017), to identify opportunities and design market strategies for the private sector (Gottfried et al., 2021) and to analyze relationship between citizens and government (Bagozi et al.,

2021). Classification and clustering algorithms, regression models and feature selection were used to predict taxpayer groups (Cha, 2020), to formulate an environmental management strategy (Kang et al., 2021), to classify government expenditure records (De Oliviera, 2021) or to make analysis of open data judgments (Metsker, 2019). For this research, knowledge extraction would be made by using Natural Language Processing (NLP). Knowledge extraction or knowledge discovery is extraction of previously unknown, and potentially useful information from data (Frawley et al., 1992). NLP models can use, for example, part-of-speech tagging, chunking and parsing to describe syntactic information or use word-sense disambiguation, semantic role labeling or named entity extraction for gaining semantic information (Collobert et al., 2011). Statistical technique used behind NLP is Probabilistic Latent Semantic Analysis (PLSA), the purpose of which is to identify or distinguish contexts of words without a need to recourse to a dictionary or thesaurus (Hofmann, 2001). In other words, we can use NLP to reveal similarities of topics by grouping together words of a common context. The use of this method for identifying the impacts of OGD could not be detected in the literature, while the use of the method for various purposes is quite common. Among other things, NLP is used to develop smart linked open government data (Sinif and Bounabat, 2019), to develop methods for classifying government documents (Peña et al., 2018; Song et al., 2019), or to predict the emergence of civic activism and protests (Kallus, 2014).

As there is not much research with the chosen methodology for the selected purpose in the literature, the same is true for research on measuring the impact of OGD. Quantifying the economic impact of open data is relatively complex, as the most important benefits are indirect (Huyer and van Knippenberg, 2020). Although the analysis of the economic benefits of open data is a good way to describe the effect of government openness, the impact of open data is not connected to economic field only, as public sector openness brings other benefits to society by increasing government or institutional responsiveness (Keserű and James Kin-sing, 2015). As Keserű & James Kin-sing (2015) further point out, evidence of the social and political impact of OGD is extremely rare. Further, Carrara et al. (2015) found out that the majority of studies conducted on the impact of OGD are preliminary assessments, which are given on the basis of classical mechanisms for obtaining feedback - surveys. According to the guidelines of the Organisation for Economic Co-operation and Development (OECD), member states of the European Union (EU) are required to submit annual surveys to review the state of open data policies

(OECD, 2018). While surveys can provide important feedback, governments face constraints on staffing and funding in the regular collection, maintenance, and exchange of data as they meet other priorities in their work (Zuiderwijk & Janssen, 2014). On the other hand, Young & Verhulst (2016) report that case studies of individual initiatives can help to better understand the impact of open data, which we see as an opportunity to relieve public sector employees and use existing use cases to identify effects in an automated manner.

2 Problem definition

The problem we are solving in this research is therefore the identification of impact areas of OGD with the help of use cases, which would be compared with the existing impact areas, defined in the OECD surveys. The outcome of this research will be validated in cooperation with Ministry of Public Administration of the Republic of Slovenia who also provided us with their surveys.

Current impact areas, identified by OECD are policy, impact, portals and data quality. These are intertwined with the identified impacts from the literature, which are mainly grouped into three categories - operational and technical, social and political, and economic (Janssen et al., 2012; Zuiderwijk et al., 2018). Problem debated in this paper is also recognized in the literature, as many authors cite monitoring of the feedback and assessment of the actual impact of OGD for further research (Attard et al., 2015; Johnson, 2016; Lourenço, 2016; Ruijer & Martinius, 2017; Wilson & Cong, 2020, Zuiderwijk et al., 2018, Zuiderwijk & Janssen, 2014).

Research questions that we aim to address are the following:

- Can we use text mining methods on open data use cases to determine an objective assessment of OGD impact by automatic extraction of semantic structure?
- Can we use the proposed methodology to validate impact areas defined by OECD?

This section of the paper can best be summed up by Janssen et al. (2012, p. 260) "The main challenge is that open data has no value in itself; it only becomes valuable when used".

3 Methodology

The methodological approach for this research fits under the Design Science Research (DSR) approach, where an IT artefact, rooted in real-world problem is designed (Hevner et al., 2004). For understanding of the data, several authors (Azevedo & Santos, 2008; Bosnjak et al., 2009; Nadali et al., 2011; Schafer et al., 2018) recommend the use of CRISP-DM methodology.

CRISP-DM aims, as described by Wirth & Hipp (2000), to make larger data mining projects reliable, faster and cost efficient. This is achieved by following six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment, nevertheless, the phases are usually intertwined and not linear (Bohanec et al., 2017).

Following CRISP-DM methodology, we will first start by analyzing OGD impact ecosystem, evaluation surveys and use cases from European Data portal (Data Europa EU, n. d.). By analyzing nearly 1000 use cases, we should be able to get a better understanding of the research area and the data. Since data understanding is the second phase of CRISP-DM we will then start by preprocessing the data. As preprocessing of the data is one of the first and most critical phases in data mining (Xiang-Wei & Yian-Fang, 2012), we suspect it will take the most effort. Next, modelling of the data will be made using text mining techniques, among which we will use NLP with its basic components such as word tokenization, stop words removal, part of speech tagging and stemming/lemmatization (Collobert et al., 2011). Finally, evaluation of impact areas with the help of Ministry of Public Administration will be made and a model for indexing, analyzing and searching heterogeneous document collections in order to extract knowledge from textual contents through NLP will be presented. A brief idea of the preliminary model is displayed in Figure 1 where keyword extraction from case studies is first made on the right side of the figure. Keywords from the documents are then grouped together and formed into collections or impact areas. Groups of variables are also created in order to better distinguish between groups of keywords and collections. Administration dashboard is the central entity, where inputs (input documents for impact assessment analysis) are uploaded and then classified based on similarity of specific impact area or impact area variables.

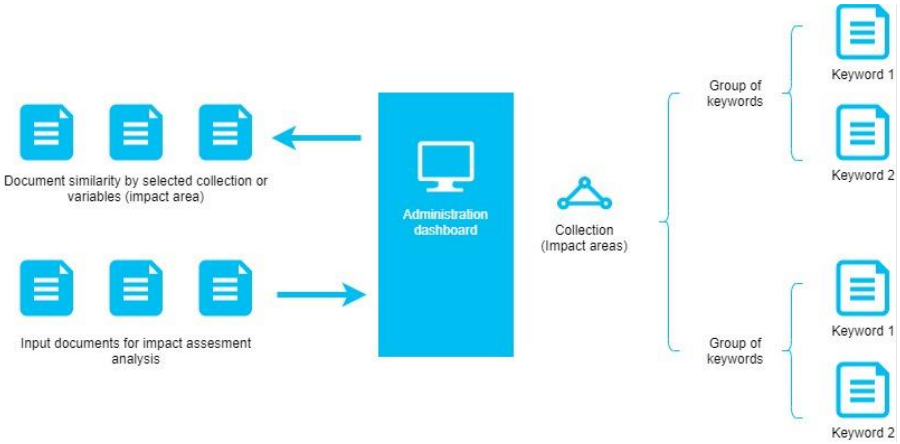

Figure 1: Preliminary schema of semantic structure for artefact development.

4 Expected results

Through our research, we aim to assess impact areas of OGD in an automated manner through use case analysis. We aim to cross-validate impact areas by comparing them with OECD surveys and further develop a model that will help governments with classification of impact areas based on provided inputs.

Based on research gaps we defined in this paper, we believe that assessing impact areas of OGD is important and will play an important role when benefits of open data initiatives are evaluated and summarized. Results of our research will contribute to better understanding of the actual impact of OGD and will help governments to prepare beneficial and focused datasets for providing even more value for citizens based on DSR's practical aspect of this research. By understanding the actual impact of OGD, more focused approach for opening the data and feedback mechanisms will be introduced from the governments. By designing a model for OGD impact assesment with the proposed methodology that hasn't been yet used for the addressed problem, we are filling up research gaps from the literature and that is essential for further academic research and for advancing knowledge in the field of open data.

References

Afful-Dadzie, E., & Afful-Dadzie, A. (2017). Liberation of public data: Exploring central themes in open government data and freedom of information research. International Journal of Information Management, 37(6), 664–672. doi:10.1016/j.ijinfomgt.2017.05.009

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government Information Quarterly, 32(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Attard, J., Orlandi, F., & Auer, S. (2016). Value Creation on Open Government Data. 2016 49th Hawaii International Conference on System Sciences (HICSS). doi:10.1109/hicss.2016.326

Azevedo, A., & Santos, M. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conf. Data Mining.

Bagozzi, B. E., Berliner, D., & Almquist, Z. W. (2019). When does open government shut? Predicting government responses to citizen information requests. Regulation & Governance. doi:10.1111/rego.12282

Bohanec, M., Robnik-Šikonja, M., & Kljajić Borštnar, M. (2017). Decision-making framework with double-loop learning through interpretable black-box machine learning models. Industrial Management & Data Systems, 117(7), in print (July, 2017b), http://dx.doi.org/10.1108/IMDS-09-2016-0409

Bosnjak, Z., Grljevic, O., & Bosnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. 2009 5th International Symposium on Applied Computational Intelligence and Informatics, 509-514. doi: 10.1109/SACI.2009.5136302

Carrara, W., Chan, W. S., Fischer, S., & van Steenbergen, E. (2015). Creating value through Open Data. European Commission. https://doi.org/10.2759/328101

Cha, T. (2020). Open Government Data for Machine Learning Tax Recommendation. The 21st Annual International Conference on Digital Government Research.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. Journal of Machine Learning.

Crusoe, J., Simonofski, A., Clarinval, A., & Gebka, E. (2019). The Impact of Impediments on Open Government Data Use: Insights from Users. 2019 13th International Conference on Research Challenges in Information Science (RCIS). doi:10.1109/rcis.2019.8877055

De Oliviera Almeida, G., Revoredo, K., Cappelli, C. & Maciel, C. (2021). Method for Improvement of Transparency: Use of Text Mining Techniques for Reclassification of Governmental Expenditures Records in Brazil. International Journal of Business Intelligence and Data Mining, 1, 1. https://doi.org/10.1504/IJBIDM.2021.112989

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. AI Magazine, 13(3), 57. https://doi.org/10.1609/aimag.v13i3.1011

Gottfried A, Hartmann C & Yates D. (2021). Mining Open Government Data for Business Intelligence Using Data Visualization: A Two-Industry Case Study. Journal of Theoretical and Applied Electronic Commerce Research, 16(4):1042-1065. https://doi.org/10.3390/jtaer16040059

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75-105. doi:10.2307/25148625

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42, 177–196. https://doi.org/10.1023/A:1007617005950

Huyer, E., & van Knippenberg, L., (2020) The Economic Impact of Open Data: Opportunities for value creation in Europe, Capgemini Invent. European, Data Portal. Doi:10.2830/63132.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258–268.

Jaeger, P. T., & Bertot, J. C. (2010). Transparency and technological change: Ensuring equal and sustained public access to government information. Government Information Quarterly, 27(4), 371–376. doi:10.1016/j.giq.2010.05.003

Johnson, P. A. (2016). Reflecting on the success of open data: How municipal government evaluates their open data programs. International Journal of E-Planning Research, 5(3), 1-12.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Linked Open Government Data Analytics. Electronic Government, 99–110. doi:10.1007/978-3-642-40358-3_9

Kallus, N. (2014). Predicting crowd behavior with big public data. Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion. doi:10.1145/2567948.2579233

Kang, A., Ren, L., Hua, C., Song, H., Dong, M., Fang, Z. & Zhu, M. (2021). Environmental management strategy in response to COVID-19 in China: Based on text mining of government open information. Science of The Total Environment, Vol. 769. https://doi.org/10.1016/j.scitotenv.2021.145158.

Keserű, J. & James Kin-sing C. (2015). The Social Impact of Open Data." Sunlight Foundation. http://www.opendataresearch.org/dl/symposium2015/odrs2015-paper20.pdf

Lourenço, R. P. (2016). Evidence of an Open Government Data Portal Impact on the Public Sphere. International Journal of Electronic Government Research, 12(3), 21–36. doi:10.4018/ijegr.2016070102

Metsker O., Trofimov E., Sikorsky S. & Kovalchuk S. (2019) Text and Data Mining Techniques in Judgment Open Data Analysis for Administrative Practice Control. Communications in Computer and Information Science, Vol 947. https://doi.org/10.1007/978-3-030-13283-5_13

Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. 2011 3rd International Conference on Electronics Computer Technology. doi:10.1109/icectech.2011.5942073

Open Government Data. (b. d.). https://www.oecd.org/gov/digital-government/open-government-data.htm

OECD (27. 9. 2018). Open Government Data Report. https://www.oecd.org/gov/open-government-data-report-9789264305847-en.htm

Peña, P., Aznar, R., Montañés, R. & Hoyo, R. (2018). Open Data for Public Administration: Exploitation and semantic organization of institutional web content. Sociedad Española para el Procesamiento del Lenguaje Natural, 61: 155-158. doi:10.26342/2018-61-21

Pereira, G. V., Macadar, M. A., Luciano, E. M., & Testa, M. G. (2016). Delivering public value through open government data initiatives in a Smart City context. Information Systems Frontiers, 19(2), 213–229. doi:10.1007/s10796-016-9673-7

Roa, H. N., Loza-Aguirre, E., & Flores, P. (2019). A Survey on the Problems Affecting the Development of Open Government Data Initiatives. 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG). https://doi.org/10.1109/icedeg.2019.8734452

Ruijer, E. H. J. M., & Martinius, E. (2017). Researching the democratic impact of open government data: A systematic literature review. Information Polity, 22(4), 233–250. doi:10.3233/ip-170413

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. Information Polity, 22(1), 1–24. https://doi.org/10.3233/ip-160012

Schafer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018). Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes. 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). doi:10.1109/itmc.2018.8691266

Sinif, L., & Bounabat, B. (2019). A general framework of smart Open Linked Government Data. Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis - ICGDA 2019. doi:10.1145/3318236.3318243

Song, Y., Li, Z., He, J., Li, Z., Fang, X., & Chen, D. (2019). Employing Auto-Annotated Data for Government Document Classification. Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence - ICIAI 2019. doi:10.1145/3319921.3319970

Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/5k46bj4f03s7-en.

Data Europa EU. (n. d.). Use Cases. Retrieved May 4, 2021, from https://data.europa.eu/en/impact-studies/use-cases

Wilson, B., & Cong, C. (2020). Beyond the Supply Side: Use and Impact of Municipal Open Data in the U.S. Telematics and Informatics, 101526. doi:10.1016/j.tele.2020.101526

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pg: 29 – 39.

Xiang-wei, L., & Yian-fang, Q. (2012). A Data Preprocessing Algorithm for Classification Model Based On Rough Sets. Physics Procedia, 25, 2025–2029. doi:10.1016/j.phpro.2012.03.345

Yan, A., & Weber, N. (2018). Mining Open Government Data Used in Scientific Research. Lecture Notes in Computer Science, 303–313. doi:10.1007/978-3-319-78105-1_34

Young, A. & Verhulst, S. (2016). The Global Impact of Open Data: Key Findings from Detailed Case Studies Around the World. O'Reilly Media.

Zuiderwijk, A. & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. Government Information Quarterly, 31(1), 17–29. doi:10.1016/j.giq.2013.04.003

Zuiderwijk, A., Shinde, R., & Janssen, M. (2018). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results?. International Review of Administrative Sciences, 85(4), 645-672.