Analysis of Selected Characteristics of Open Data Inception Portals in the Context of Smart Cities IoT Data Accessibility

Paweł Dymora[1], Mirosław Mazurek[2] and Bartosz Kowal[3]
Faculty of Electrical and Computer Engineering, Rzeszów University of Technology, al. Powstańców Warszawy 12, 35 959 Rzeszów, Poland

Keywords: Open Data, Open Government, Knowledge Management, Complex Systems, Data Mining, IoT, Smart City.

Abstract: In this study, we focus on Open Government Data, which is the sphere of public services where such type of data can be useful. In the Industry 4.0 concept, the primary data source is the IoT infrastructure. Open Data is of considerable importance for the software development process. The issue of Open Data is becoming a significant challenge nowadays. Especially when it comes to preparing data for sharing, analyzing it, and searching for hidden dependencies, which opens up new possibilities for computing and artificial intelligence. The paper shows that the architecture of solutions existing, e.g., in Poland, follows global trends. Together with statistics based on the Socrata portal, it can be noticed that these data can be and are successfully used for data processing. New methods and software are being developed for processing data as we write. The vast majority of software is data-driven, and data are needed for verification and validation. The article presents a comprehensive analysis of available open data portals with data.json files as also the analysis of the most commonly used data formats for Open Data Network portal databases.

1 INTRODUCTION

This study is on Open Data in the context of Open Governments. Open Data has become one of the critical transformational economic forces of the XXIst century. They should no longer be associated solely with the Internet. When the algorithm recommends a book in a bookstore or completes a phrase in a search engine, solutions related to Big Data are involved. The same can be said in the idea of Industry 4.0 when the continuous monitoring of production processes is required (Janssen et al., 2017). The concept of a smart city defines a city that uses data and technology to improve the lives of residents and companies that live there. The critical technology for the success of smart city initiatives, regardless of whether pollution levels or road conditions are improving, is the Internet of Things. IoT is a network of physically connected devices, such as vehicles or home appliances, that enable these "things" to connect and exchange data. This, in turn, creates unique physical and digital connectivity options - through data analysis (Open data system) - to improve performance (in both the public and private sectors), provide economic benefits, and improve living conditions. IoT is also a concept of building telecommunication networks and IT systems with a high degree of dispersion. Such systems can serve, among others, the creation of intelligent control and measurement systems, analytical systems, or control systems, practically in every area of life, economy, or science. IoT is a concept of IT architecture, which enables cooperation (interoperability) of various ICT systems supporting various field applications. It is based on software for data exchange, their processing, system management, and its protection and sharing. IoT sensor infrastructure is used for many complex functions, generating data and/or controlled via the Internet, autonomous transport devices, industrial robots, mobile devices (smartphones, smart-watches, etc.), intelligent digital metering devices.

According to (Fischer et al., 2015), *"Between 2016 and 2020, the market size of open data is*

[1] https://orcid.org/0000-0002-4473-823X
[2] https://orcid.org/0000-0002-4366-1701
[3] https://orcid.org/0000-0002-7909-6484

Dymora, P., Mazurek, M. and Kowal, B.
Analysis of Selected Characteristics of Open Data Inception Portals in the Context of Smart Cities IoT Data Accessibility.
DOI: 10.5220/0010117600670074
In *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020)*, pages 67-74
ISBN: 978-989-758-478-7

expected to increase by 36.9%, to a value of 75.7 bln EUR in 2020. The forecasted number of direct open data jobs in 2016 is 75,000 jobs. From 2016 to 2020, almost 25,000 extra direct open data jobs are created. The forecasted public sector cost savings for the EU28+ in 2020 are 1.7 bln EUR. Efficiency gains are measured by a qualitative approach. A combination of insights around efficiency gains of open data and real-life examples are provided. To measure the success of open data policies, a series of recommendations is put forward to help governments keep track of the direct and indirect benefits of their policies. This is key in further accelerating the publishing of open data and encouraging its re-use". The presented study, analysis the amount of data available for individual Open Data Network and open Data Inception platforms in the context of the most commonly used data formats for Open Data Network portal databases.

2 RELATED WORK, CONCEPTS, CHALLENGES OF OPEN DATA

For many years people were lived in a world of *"small data samples."* However, nowadays, we use more and more often, big data sets that improve our perception of reality (Śniegocki et al., 2014).

Many economic sectors are generating vast amounts of data: companies working with e-content (websites, e-books, music, videos), monitoring of Internet users' activity, social media, digitization of money with billions of e-transactions, the healthcare sector, etc. Data produced by monitoring patients and the effects of different treatments meet the 3Vs criteria (Śniegocki et al., 2014; Gil et al., 2017). The 3V model presents the most essential and necessary features of the big data type. Volume is a feature that determines the size of collected and stored data. The storage, management, and analysis of such a collection exceed the capabilities of standard database tools. Usually, these are numbers reaching at least a few terabytes. The large corporations process data even of several dozen petabytes or even exabytes. Another feature is Velocity – the speed of data inflow. Data appear very dynamically, practically in real-time. This means that al-most in the same second, in which information is generated, it is possible to use it. The last property is Variety regarding data sources. The data come from many, often surprising sources. They are not homogeneous – they may be text, sound or graphics files (Śniegocki et al., 2014; Gil D et al. 2017).

According to (Perera et al., 2017; Sánchez et al. 2017; Naranjo et al. 2018) states that the Internet of Things (IoT) or the Internet of Everything (IoE) aims to connect billions of intelligent objects to the Internet, which can improve the services of smart cities. As stated in (Harmon et al., 2015), *"the emerging Internet of Things (IoT) model is the foundation for the development of smart cities."*

Authors in (Gil et al., 2017; Dong et al., 2017) state that open data has the potential to change the way that government, citizens, and organizations exchange, access, and use data. Investigations are becoming costly when data-mining technologies are utilized. Edwards in (Edwards et al., 2015) claims that: *"Such technologies must be well designed and rigorously ground-ed, yet no survey of the online data-mining literature exists which examines their techniques, applications, and rigor."*

The concept of an Open Government is closely related to the technological changes of the last two decades, induced by the Internet. In terms of values or fundamental rights, open government and open data are independent of the technologies used to implement them. The idea of the Open Government refers to the value of the right to information and to know. Both of them are related respectively to the two fundamental objectives of implementing the open government model, which are the transparency of government activities and the involvement of citizens in them. According to (Jachowicz et al., 2013), the Open Government is: *"A new way of organizing activities in a country that uses digital technology and communication tools to increase the participation of citizens in government and to use their knowledge and commitment to solving problems more effectively."*

The critical values of an open government definition are transparency (clarity), participation, and cooperation. However, two extra values to this list can be added: efficiency and openness. Efficiency is the value of the reform of public administration. The openness is treated as the foundation of other values and a component of all activities for the benefit of an open government.

The goal of the open government is to make widely available public information resources. Ensuring openness is one of the fundamental objectives of the open government. Several types of such resources can be distinguished. Usually, resources are defined by the general term of public information or public sector information.

We can distinguish the following resources (Hofmokl et al., 2012):

- official documents and materials that are traditionally the primary form of public

information (for example, the content of legal acts, but also reports, expert opinions, and other studies);
- information on the governance processes undertaken based on these documents (e.g., the legislative process or the results of the voting in the parliament);
- raw data collected or generated by the administration, which may be treated as the basis for two types of information mentioned above.

The concept of public information resources implies access to resources. However, in the literature, the terms "public information" and "public data" are often used interchangeably today, along with the widely used term "open data." The Polish Normative Act on Access to Public Information regulates information about the internal and foreign policy, public entities, and rules of their functioning, public data, and public assets. The openness of the government requires the availability of all these types of information (Polish Journal of Laws, 2001).

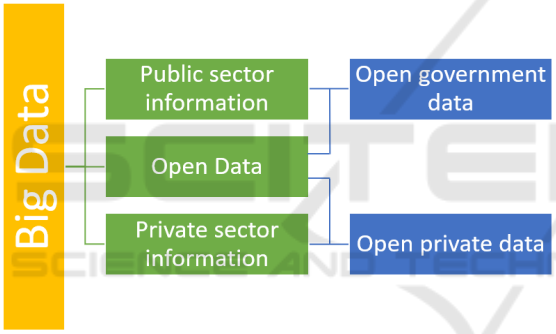

Figure 1: Big data interrelation in 5 segments.

Open Data is a subset of the commonly used term Big Data, which is a popular term to describe any collection of large information sets. The big data interrelation is shown in Fig. 1. The critical principle of Open Data philosophy is openness by default. Instead of finding the reason why the given piece of data should be open for the public to use and re-use, data owners follow the default notion that data should be open unless there are important reasons to restrict access to them (as stipulated in (Śniegocki et al., 2014)). Another essential attribute, which is often overlooked in the open data approach, is the users' ability to verify the authenticity of its source and its integrity without external alterations to the originally released data. For example, the use of electronic signatures should not be required.

Providing government data to the general public may be seen as the equivalent of open government itself, but there are new problems with state-owned data, such as trust in the way the government handles data, in particular, the identity of individuals. It is essential to consider an aspect that provides adequate guarantees regarding the ethical and moral use of data stored by the (Schauppenlehner & Muhar, 2018).

As indicated in (Schauppenlehner & Muhar, 2018), it can be seen, for example, throughout Europe, that public and local authorities provide free access to various data (e.g., statistical data, geodetic data). End-users (the general public, interest groups, students, other bodies, etc.) have free access to data, but this option requires specific knowledge, methods, and guidance to identify and use relevant content. This is done thanks to the concept of metadata.

3 ANALYSIS OF THE OPEN DATA PORTALS GATHERED BY OPEN DATA NETWORK AND OPEN DATA INCEPTION

There are specific models to represent Open Data Systems. UML diagrams are used for it. However, a context diagram is of great help. At the moment, many countries have their own open data portal in government domains. Portal Open Data Inception developed by OpenDataSoft (Open Data Inception, 2019) gathered with one portal over 2600 different open data portals. In Fig. 2 we can see that most Opendata portals are located in Europe (more than 1200 portals). In Poland, the most prominent Open data portal is Otwarte Dane, for Germany is GovData and Socrata as also data.gov in USA. On the Internet, there is also an open portal within the European Union containing open data from Member States (EU Open Data Portal).

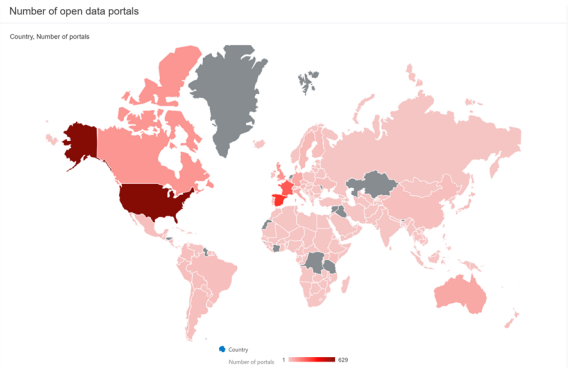

Figure 2: Location map of open data portals posted in Open Data Inception (Open Data Inception, 2019).

One of the service providers for viewing Open Data from the territory of the United States is Socrata (Socrata, 2019). This platform was created to facilitate the generation and use of resources generated by offices and cities. At the moment Open Data Network portal powered by Socrata is showing collections from 240 Opendata portals in the USA.

In our study, we conducted an analysis of the amount of data available for individual Open Data Network and Open Data Inception platforms. In order to better compare platforms and the benefits of their use, we have limited the location of suppliers to the USA, and the time period is set to 2018-2019. We examined the number of suppliers and the data available on both portals. Next, we examined the overall degree of data openness in the top 10 random portals and what type of data formats are generated in 10 top random portals.

This paper is based on the gathered data that the number of data sets shared by all Open Data Networks (Open Data Network, 2019) portals for the period of May 2018 to December 2019 was checked. We can see that a significant proportion of portals (66.4%) have no more than 100 data sets. A much smaller group are portals publishing from 101-200 (11.6%), 201-400 (11.2%) data sets. The smallest group are huge portals with over 5000 data resources. Looking at the number of published data (Fig. 3), we can see that in January 2019, 2 large portals were publishing over 3000 data sets. It is worth considering that the number of data sets can dynamically increase or decrease; some sets can be combined with each other or archived.

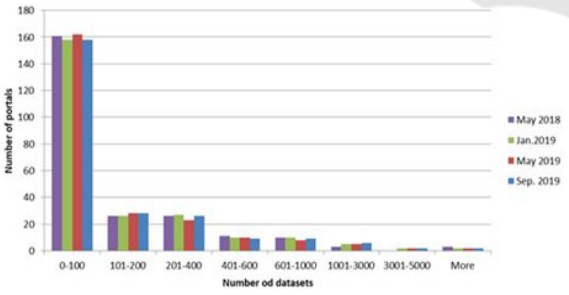

Figure 3: The number of available datasets on Open Data Network portals between May 2018-September 2019 with the usage of API.

Compared to the Open data Networks Portal, the Open Data Inception Portal is not created to display a full list of datasets. It is supposed to gather a list of all open data portals in the whole world narrowing the criteria for selecting Open Data Inception portals, only to the US country, and to compare the number of portals available on the Open Data Networks portal.

Figure 4 shows that only 64 portals (26.5%) located on the Open Data Networks portal are also on the Open Data Inceptions portal, e.g., Austin's Open Data Portal. These portals are only a small part of the Inception project (9,7%). So it can be concluded that not all portals that are on the Open Data Networks portal are on the Inception portal.

However, looking at the distribution of organizations providing data (Fig. 5), we can see that the largest part of the portals (70%) is grouped by the United Nations. 5% of the portals are grouped by the United Nations (UNESCO, Unicef, World Bank, US Databases, etc.). The second-largest group is organizations associating less than three portals (21%). These are mainly open city portals, e.g., the City of Las Vegas.

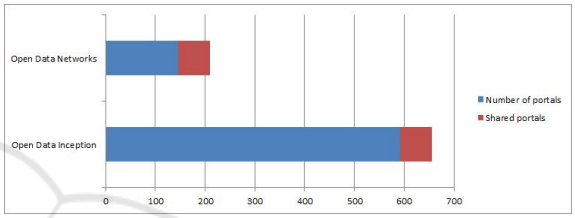

Figure 4: Shared portals between the Open Data project.

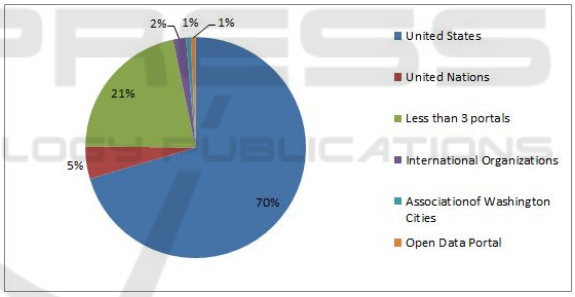

Figure 5: Open Data Inception, filter for the organization's portal in the US.

Using the API created for the Open Data Networks portal, in the period February 2018 - August 2019, we tested data formats on a daily basis. It should be emphasized that one set of data can have several different file formats. In our study, we show how many percent of formats apply to the whole number of data sets. Table 1 summarizes the data formats in individual months. As can be seen (Fig. 6), in the period February 2018 - August 2018, a significant part of the available data, as much as 74-79%, was in the octet-stream format. These could be binary files, files without extensions, executables. The situation changed in December 2018, where such a format was already 1-4%. The accessible CSV format in this period accounted for about 50 ± 3% of data resource files. There were no significant changes

in the number of published data sets. At the beginning of February 2018, the other accessible XML format accounted for 51% of the data file format. With the subsequent months, its use dropped from 51% to 43%. Both CSV and XML formats are the most commonly used formats in open data. In comparison to several data formats published in (Dymora et al., 2018), it can be seen that there was an increase of pdf (5%) as also HTML (70%) formats and a decrease of octet-stream format to nearly 1%.

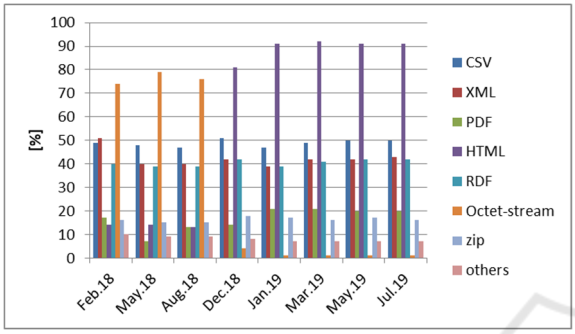

Figure 6: Chart of data formats on Open Data Networks portals.

Table 1: Datasets formats for Open Data Network portals.

	Feb.18	May.18	Aug.18	Dec.18	Jan.19	Mar.19	May.19	Jul.19
CSV	49%	48%	47%	51%	47%	49%	50%	50%
XML	51%	40%	40%	42%	39%	42%	42%	43%
PDF	17%	7%	13%	14%	21%	21%	20%	20%
HTML	14%	14%	13%	81%	91%	92%	91%	91%
RDF	40%	39%	39%	42%	39%	41%	42%	42%
Octet-stream	74%	79%	76%	4%	1%	1%	1%	1%
Zip	16%	15%	15%	18%	17%	16%	17%	16%
Others	10%	9%	9%	8%	7%	7%	7%	7%

4 OPEN DATA IN POLAND

As mentioned in the introduction section, Open Data is an essential subset of public sector information. It may include, for example, data from urban sensors, public procurement data, or health data. The state collects a wide variety of diverse data using the actions of public entities. In terms of data openness, it is vital to rediscover the value of data. It is also essential to promote innovation, mutual benefits such as getting feedback, increase transparency, and network effects.

Open Government Data in Poland is at a very early stage of development. In the EU, the first Public Sector Information (PSI) Directive was issued in 2003, but its implementation in Poland was heavily delayed. Substantial funding will be allocated to open government data projects as part of the EU's Digital Agenda 2020 and its national implementation program "Digital Poland Operational Programme" for the period between 2014 and 2020 (Open Government Data Review of Poland, 2015).

According to the Polish Normative Act on Access to Public Information and two additional regulations - Regulation of the Council of Ministers regarding the Central Public Information Repository and the Regulation of the Ministry of Administration and Digitization concerning the information resource to be made available in the Central Repository of Public Information (Polish acronym – CRIP) was created (Polish Journal of Laws, 2001). CRIP is an IT model that facilitates the access and re-use of information resources of particular importance for the development of innovation, contributing to the progress of information society as well as the resources of public information.

Using the API created for the Otwarte Dane (Otwarte dane, 2019), in the period of February 2018 - December 2019, we tested the amount published data by each institution in the Polish portal and the number of people visiting institutions for the search of datasets. As can be seen (Fig. 7), we divided the data into three periods February 2018, December 2018, and December 2019. ZUS (from Polish: Social Security Institution) publishes the most data sets (73 sets in 2018, 91 sets in 2019, followed by GUS (from Polish: Central Statistical Office) 73 in 2018 and 84 sets in 2019). Despite a large number of published data sets, ZUS is not often visited (22,000 visits in a year). As can be seen (Fig. 8), the most frequently visited institution is the Ministry of Digitization (a total of over 330,000 visits, the next one is GUS 203,000).

The openness of data sets on the Polish Open Data portal is based on a five-star system for implementing open data proposed by Tim Berners-Lee. The higher the level of data openness, the better the data is prepared for further processing and processing of the data contained therein. All open data is shared without any restrictions for any commercial and non-commercial purposes. Some of the files require informing the unit was providing data about the type and purpose of data processing. Good practice in creating data is publishing data, at least at the second level of openness (Otwarte dane, 2019).

First-level data is usually easy to publish files such as PDF, jpg format. Making them available on the Open Data portal does not require additional administrator's work by providing data to the technical conditions of the formats. In the second

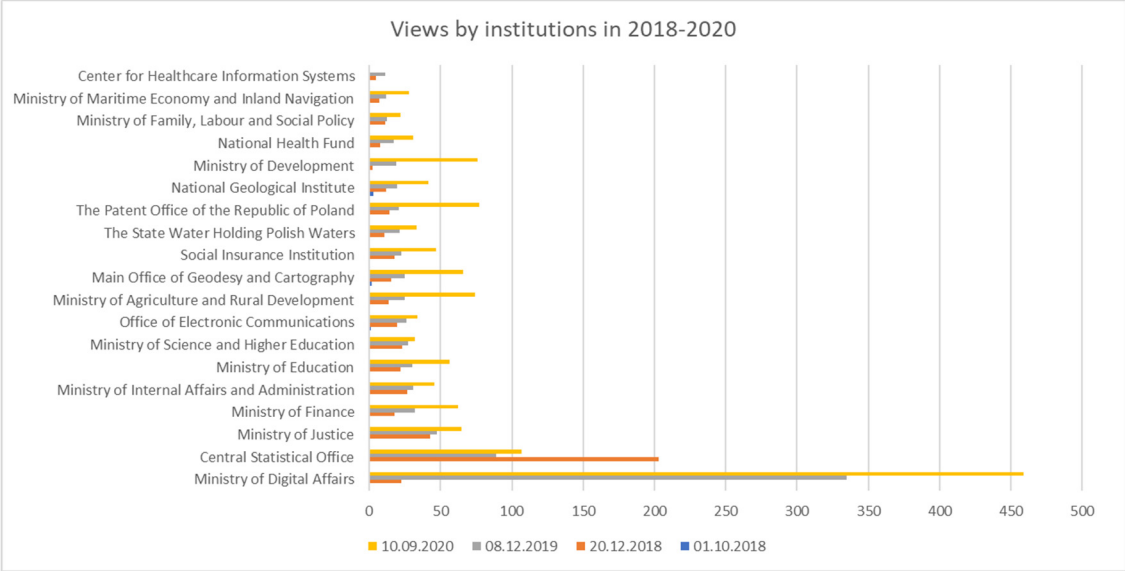

Figure 7: The number of datasets published by the institution in 2018-2020.

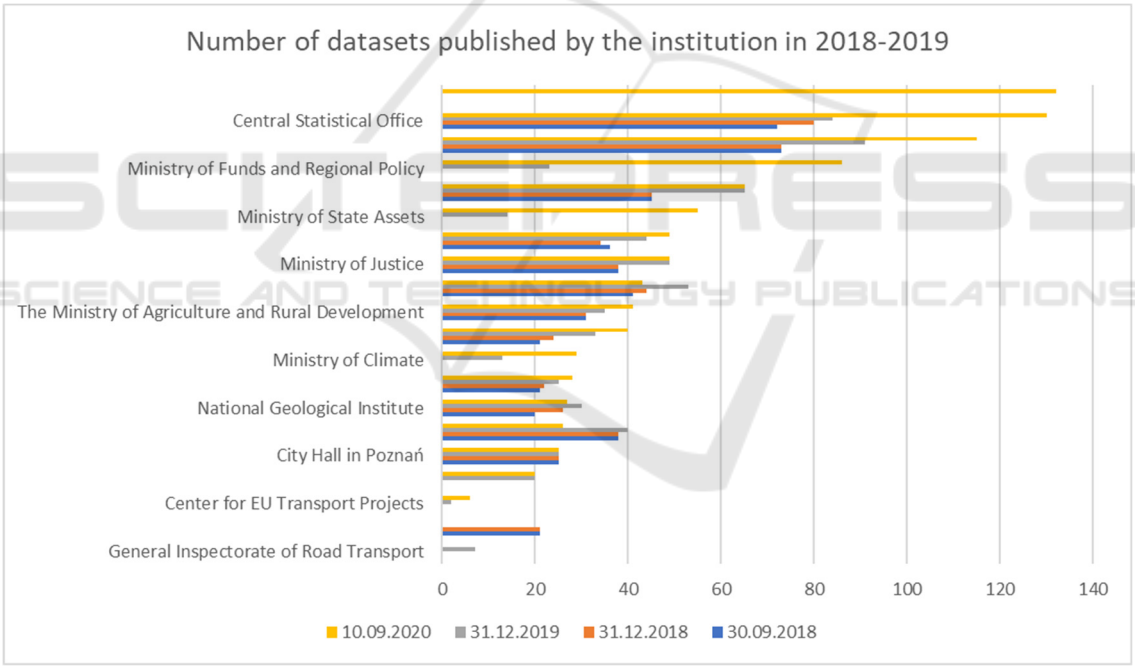

Figure 8: Views by institutions in 2018-2020.

level, by structuring the document, the user can export the file to another structured format. A specific format (e.g., Excel table) allows calculations and visualization of data contained in the document. In the third level, the user has the option of editing the downloaded file for their own needs, without the need for additional software, e.g., CSV data.

From the administrator's point of view, the file is as quick and straightforward to publish as the previous two. In the fourth level, the data must be able to share them via a URL link to other services automatically. Thanks to the possibility of linking the document, the user can share the document, e.g., on social media or his website. It also can save the document in the tabs of the browser for quick access to data. The fifth, highest level is the data presented in the list of other, current, open data sets. Correlated

data make it easier to search for the desired, and the user receives other related files in the form of hints.

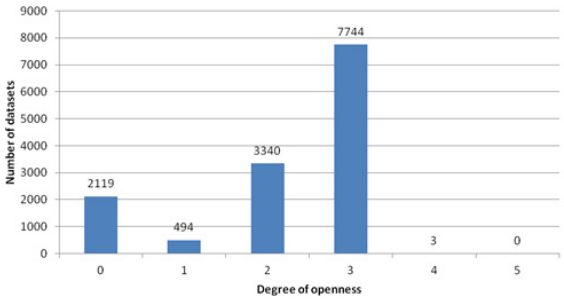

Figure 9: Degree of openness on the Polish Open Data portal.

Fig. 9 shows the level of data openness on the Polish portal. There is no tier 5 data on Polish Open Data, probably due to the existence of only one portal and a reason-ably young portal (less than three years of existence). The most significant part of published data is in level 3; more than half of the data is saved in editing formats, e.g., CSV. It is worth noting that there is a level 0 in this portal, which is treated as the default level if the organization has not set it.

5 CONCLUSIONS

Nowadays, Open Data becomes a significant challenge. Their analysis and looking for hidden dependencies open new possibilities for computational and artificial intelligence. Open data may include, for example, data from city sensors, public procurement data, or health data. The state collects a wide variety of data using public entities. The open data has been growing for several years, and propagating and popularizing solutions will allow better use of information. The organizations are committed to greater transparency, participation and greater cooperation with the population, businesses and research communities. Local governments have released data about their finances and operations in the interest of the good government and citizen participation. As a result, numerous software development environments and tools are available for different platforms. They make these open data more accessible, useful, and comparable.

Detailed analysis showed that the openness of data sets in the Polish Open Data portal is based on a five-star system of implementing open data. As has been shown, the architecture of solutions existing, for example, in Poland is following worldwide trends. Together with statistics based on Socrata portal, it can be seen that these data can be and are successfully used for data processing. This ensures that the level of data openness ensures their rapid further processing. Thanks to this, as in global portals, data can be shared without any restrictions for commercial and non-commercial purposes.

REFERENCES

Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using big and open linked data (BOLD). Information Systems Frontiers, 19 (2), pp. 189–195, https://doi.org/10.1007/s10796-017-9746-2

Fischer, S., Van Steenbergen, E., Carrara, W., & Chan W. S. (2015). Creating Value through Open Data. 1 (2015), pp. 1–112. Issue 1, https://doi.org/10.2759/328101

Śniegocki, A., Buchholtz, S., Bukowski, M. (2014). Big and Open Data in Europe: A growth engine or a missed opportunity? Demos EUROPA & WISE Institute. Vol. 1, Issue 1, pp. 1 –116.

Gil, D., Song, I-Y., Aldana, J. F., & Trujillo, J. (2017). Big Data. New approaches of model-ling and management. Computer Standards & Interfaces, 54, pp. 61–63. https://doi.org/10.1016/j.csi.2017.03.006

Perera, C., Qin, Y., Estrella J. C., Reiff-Marganiec S., & Vasilakos A. V. (2017). Fog Computing for Sustainable Smart Cities. Comput. Surveys 50, pp. 1–43, https://doi.org/10.1145/3057266

Sánchez, H., González-Contreras, C., Agudo, J.E., & Macías, M. (2017). IoT and iTV for Interconnection, Monitoring, and Automation of Common Areas of Residents. Appl. Sci. 2017, 7, 696.

Naranjo, P.G.V., Pooranian, Z., Shojafar, M., Conti, M. & Buyya, R. (2018). FOCAN: A fog-supported smart city network architecture for management of applications in the internet of everything environments. Journal of Parallel and Distributed Computing, (JPDC), ISSN: 0743-7315

Harmon, R. R., Castro-Leon E, & Bhide, S. (2015). Smart cities and the Internet of Things, 2015 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, pp. 485-494.

Dong, H. W., Singh, G., Attri A., & El Saddik A. (2017). Open Data-Set of Seven Canadian Cities. IEEE Access, 5, pp. 529–543,

Edwards, M., Rashid, A., & Rayson P. (2015). A Systematic Survey of Online Data Mining Technology Intended for Law Enforcement. Comput. Surveys, 48 (1), pp. 1–54, https://doi.org/10.1145/2811403

Jachowicz, Ł., Młynarski G., & Tarkowski A. (2013). Otwarty rząd w Polsce. Kulisy programu Opengov, Retrieved from https://centrumcyfrowe.pl/wp-content/uploads/2013/06/Otwarty-rzad-w-Polsce-Publikacja-OPENGOV-v1-0.pdfSmith, J., 1998. *The book*, The publishing company. London, 2nd edition.

Hofmokl, J. et al. (2012). Mapa drogowa otwartego rządu w Polsce. Retrieved from https://ngoteka.pl/

Polish Journal of Laws. (2001). Polish Normative Act on Access to Public Information, Retrieved from http://unpan1.un.org/intradoc/groups/public/documents/unpan/unpan034035.pdf
Socrata (2019). Retrieved from https://www.tylertech.com/products/socrata
Schauppenlehner, T., Muhar, A. (2018). Theoretical Availability versus Practical Accessibility: The Critical Role of Metadata Management in Open Data Portals. Sustainability 10, 545
Open Data Inception - 2600+ Open Data Portals Around the World (2019). Retrieved from https://opendatainception.io
Open Data Network (2019). Retrieved from https://www.opendatanetwork.com
Dymora, P., Mazurek, M., Kowal, B. Open data - an introduction to the issue. ITM Web of Conferences. 2018, vol. 21, pp. 1-13
Open Government Data Review of Poland. OECD Digital Government Studies (2015), https://doi.org/10.1787/9789264241787-en
Polish Journal of Laws. (2001). Polish Normative Act on Access to Public Information, Retrieved from http://unpan1.un.org/intradoc/groups/public/documents/unpan/unpan034035.pdf
Otwarte dane (2019). Retrieved from https://dane.gov.pl.