Open data – an introduction to the issue

Paweł Dymora[1,*], and *Mirosław* Mazurek[1], and *Bartosz* Kowal[1]

[1] Rzeszów University of Technology, Department of Complex Systems, Al. Powstancow Warszawy 12, 35-959 Rzeszów, Poland

Abstract. Rapidly developing of internet technologies and digitalization of government generate more and more data. Databases from various public institutions and private sectors, e.g. in the fields of economics, transport, environment and public safety are publishing in the global Internet network, so that any user can browse them without additional charges. Most of this data is published on the open data portals. Open data - that is, "open", public data can allow the processing and analysis of information contained in them completely free of charge. This article is an introduction to a fairly new area of issues such as "open data" or "open government", presents the main mechanisms of accessing to data in public open data portals and also propose a conceptual open data/government model.

1 Open Data

The term of “open data” was best defined by the Open Knowledge International non-profit organization [1]. It assumes that open data means the free data that can be freely used, re-used and distributed by anyone for any purpose [2]. This is a somewhat idealized definition of open data. Most published databases do not have all the listed properties [2 - 4]:

- *Free* - access to data must be free,
- *Availability* - data must be available as a whole without any intentional errors. These data must also be available in a convenient and modifiable form, e.g. CSV, JSON, GeoJSON, KML, XML, RSS or RDF Turtle,
- *Re-use and redistribution* - the data provider must allow work on the contents of the database and the ability to publish these data with other databases,
- *Use of open data* - every user has the opportunity to work on data, reuse and redistribute them regardless of their purpose.

A large part of the available open data is published by government institutions. The most of them publish in categories [5]:

- social issues, law, political life,
- science, education and communication,
- geography, environment,
- economic life, agri-food sector,
- employment and work,

* Corresponding author: dymora.pawel@prz.edu.pl

- finance and trade,
- energy, industry, production, technology and research,
- agriculture, forestry and fishing,
- European Union, international relations, international organizations,
- transport, enterprise and competition.

1.1 Open Government Data

Open Government Data (OGD) is a social initiative that refers to the public disclosure of information created by public offices. Most public offices collect large amounts of data during routine activities. The main purpose of OGD is to allow free reading of data to any user who would be interested in such information. This initiative is based on the same principle as open data. Data must be available free of charge and access to data has to be unlimited and one can use and redistribute data [7].

The governments of many countries, not only the European Union, have introduced the OGD concept aimed at providing the citizens of the country with access to the collected data by public institutions. The main purpose is often to enhance the knowledge of citizens on how public administration works. It is worth to emphasize that actually most of the data collected by offices is not published because the law does not allow the publication of certain data. Most of them are covered by various security clauses. OGD is to enable residents to analyze data and consequently allows better decisions, suggestions regarding the functioning of a given unit, or detection of irregularities [6 -7].

Referring to the OGD, in the European Union from 2012 the open data portal was created to collect data from the public sector of the EU member states. All collected data is in accordance with the terms of open data. Only a small portion of the data is subject to the conditions for their use and processing. The initiative to create a portal with the open data concept was the stimulation and economic development of the each of the EU member states, as well as the European Union itself. The main objective was to improve the transparency of operation of the state institutions that generated a large amount of data. The advantage of entering open data was the possibility of using them, e.g. in applications promoting a given city, increasing the city's reputation [8 -11]. Depending on the base category, the update time of the records varies. In the case of the finance category, specifically the budget arrears database, the updates are carried out every six months, while the agricultural information, for example about fruit and vegetable prices, is updated every month [6, 11 - 13].

1.2 Socrata - open data portal

Open data are placed on the Internet. According to the dataportals.org service [5], over 524 different open data portals have been created. Most of these types of databases can be found on specially designed portals. For Europe, the largest portal providing open data is the European Union Open Data Portal (EU ODP) containing the official open databases from all of its member countries [10]. The Polish equivalent of EU ODP is a “dane publiczne” portal [6], containing data only from the territory of Poland. It is worth mentioning the Socrata portal [14], which stores existing official data, mainly from the United States. It enables easy downloading of data from the portal using specially written libraries for various applications, e.g. the Socrata package for the language R.

Socrata is already mentioned as one of the providers of open data browsing services. Socrata is a widely used government solution to work on open data. It provides a data platform and services in the cloud for government organizations in the United States, i.e. urban, state, federal, etc. The advantage of using this portal is automation of the public

sector data flow to a commonly accessible website so that the interested people can easily find and use government data. Socrata has developed the application programming interface for open data (API). Access to public data in their database can be obtained through a common, standards-based application interface. Socrata allows to check the type of database columns and data stored in them. It has ready-to-use functions for downloading data for various applications and environments, e.g. PhpSoda, .NET, Java, PHP, Python, Ruby, R and many others [14].

There are two methods of downloading data from the portal without the specialized programming or analytical environments. The first one allows to write to the disk the database in a format: CSV, JSON, RDF, RSS, TSV and XML. Table 1. presents example of a part of Active Trademark Registrations database saved in .csv format and Fig. 1 shows the data according to the SoDA standard. The disadvantage of using this method is the necessity of manually downloading updates every time [14].

Table 1. Example of a part of Active Trademark Registrations database saved in .csv format [7]

Registration Number	Registration Date	Trademark Description	Correspondent Name	Address
102	03/15/1968	"SR" MONOGRAM	SUNRIVER RESORT LIMITED PARTNERSHIP	11777 SAN VICENTE STE 900
103	03/15/1968	"SUNRIVER"	SUNRIVER RESORT LIMITED PARTNERSHIP	11777 SAN VICENTE STE 900
158	01/24/1969	THREE TENNIS BALLS AND FOUR TENNIS RACQUETS	WEST HILLS RACQUET CLUB	----------
3957	04/21/1936	"CORNING"	MICHELE N KEEFER-MEHLENBACHER	CORNING INCORPORATED
4020	08/29/1936	"MOLY-KROME"	PACIFIC MACHINERY & TOOL STEEL CO	3445 NW LUZON ST

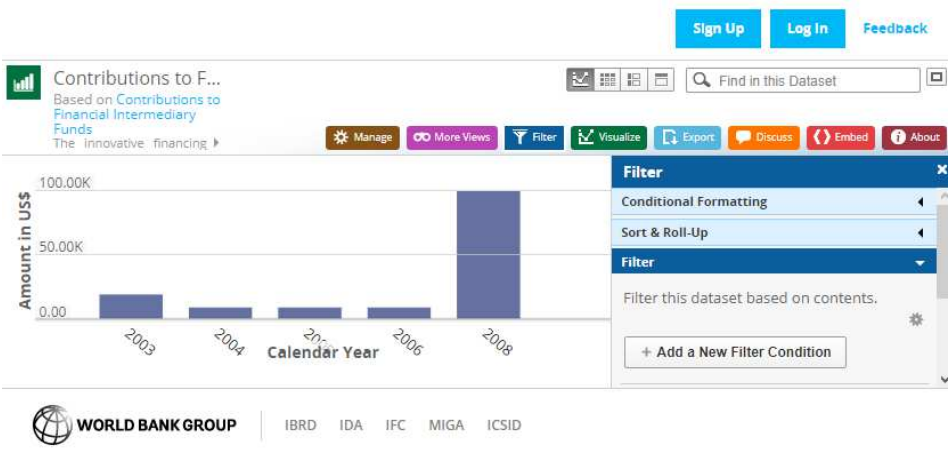


Fig. 1. Graphical presentation of data according to the SoDA standard [15].

The second method of collecting data is API. Any unauthorized user can send a small number of requests to databases that come from one shared address pool within an hour. Own API allows to make up to 1000 requests per hour, allowing for dynamic operation over data [14].

1.3 Open databases - statistics of selected portals

There is a very large number of open data databases in the Internet. To facilitate browsing and selecting the right database in the Socrata format, the data.json standard is used, containing entries for each database in the portal. From it one can read the name of the data set, the format of the record and whether the database is public.

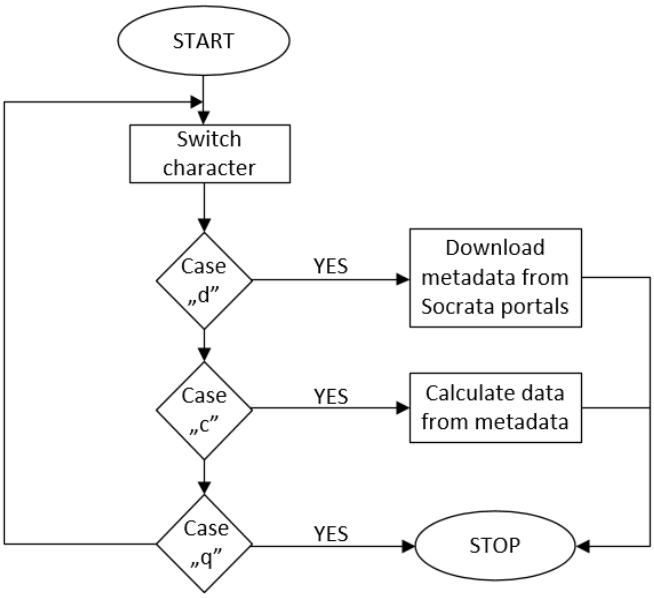


Fig. 2. Block diagram of a script that processes data from available Socrata portals.

The study analyzed selected features of Socrata's portals. A script was created to load a list of all available Socrata portals located on the opendatanetwork.com portal and its metadata. Fig. 2 presents the diagram describing the prepared script. There are currently 241 portals available. There is a possibility to download metadata about portal data sets containing, for example, information about the purpose of databases, date of creation and last modification, types of formats, etc. An additional option in the script is the ability to calculate the number of data sets and data formats that individual portals offer.

Each database may have several data formats at the same time, e.g. csv, rdf, PDF, etc. Fig. 3 shows that more than half of the databases come from government (gov) or state (edu) websites. Depending on the specificity of the portal, the number of open databases it collects is different (from 2 to 397 different sets of data).

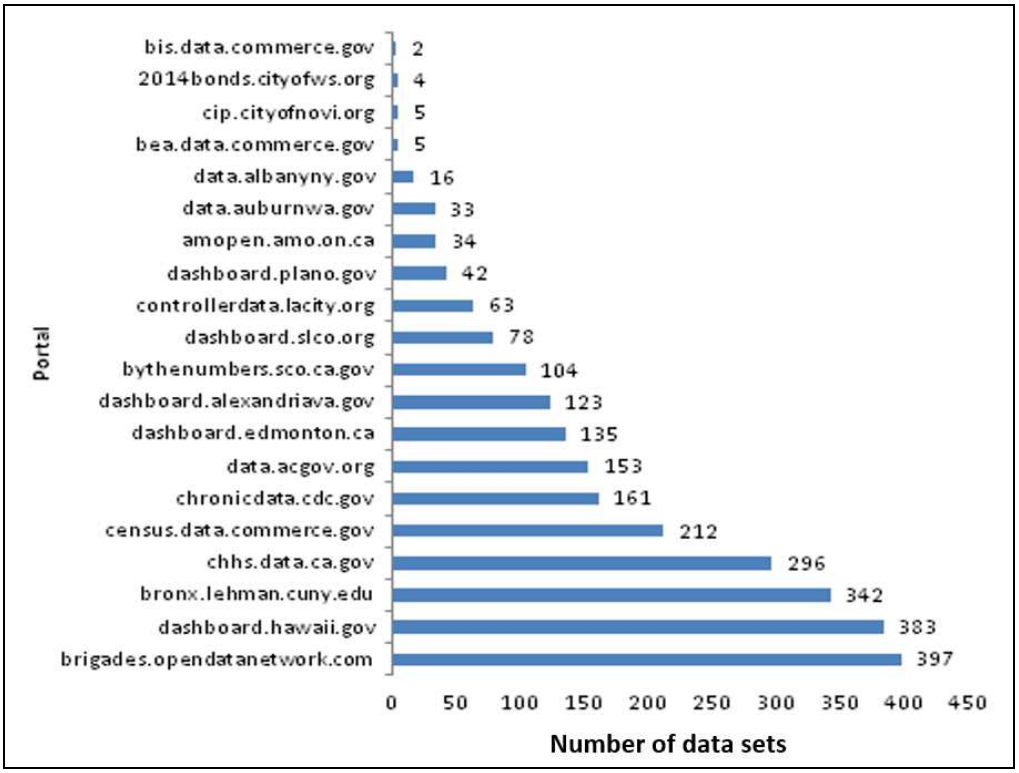


Fig. 3. Number of datasets provided by the first 20 Socrata portals based on files data.json (as of 30/06/2018) [14].

Analyzing the data.json file in terms of data formats, it can be seen that the most frequently used format is csv (49.90%), rdf (39.4%) and xml (41.5,%). The less frequently used data formats are PDF (12.1%), html (13.9%), octet stream (32.5%), vnd.ms-excel (0.2%) or zip (14.4%). The CSV type (comma delimited) is a very common format that allows to import data from a file into one's own API or your own database.

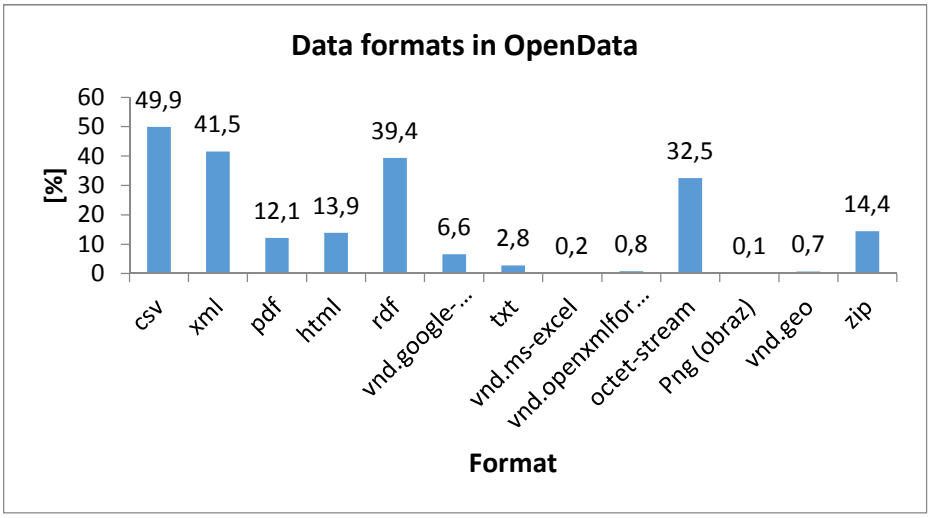


Fig. 4. The most commonly used data formats of Socrata portal databases [14].

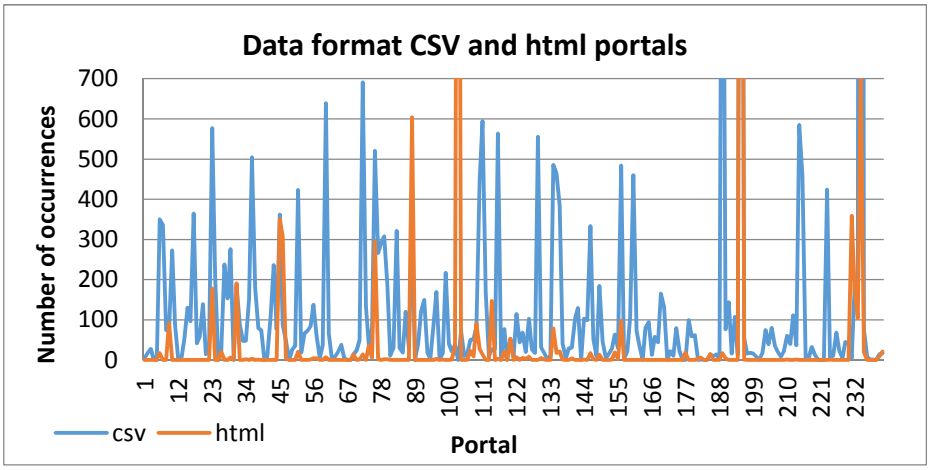


Fig. 5. Comparison of the frequency of using CSV and html data formats on Socrata portals.

Table 2 shows the number of open databases provided by various global portals, e.g. Socrata or the European EU Open Data Portal. The Socrata portal allows access to a large amount of open data, mainly from the United States (over 69,000 different sets of data). The main portal in the European Union is the EU Open Data Portal, containing over 11,000. databases from Member States' portals. Table 3 shows the number of open databases provided by various Polish portals, e.g. Local Data Bank or Local Tourist Organizations in Poland.

Table 2. Statistics of various open databases of Open Data portals (as at 30/04/2018).

Portal (name of the portal)	Number of databases
Socrata	91210
Dallas Open Data	78
New York Open Data	1430
Oregon Open Data	540
EU Open Data Portal	12313
Polskie Dane Publiczne	905
London Datastore	706

Table 3. Selected attributes of the Polish Public Data portal (as at 30/04/2018).

The name of the database	Availability	The number of columns	Time update	Average number of record changes	Recording format
Reports of the Insurance Ombudsman	Public	Not specified	Annual	All document	PDF
Local Tourist Organizations in Poland	Public with specific conditions	22	If necessary	Not specified	CSV
List of inventions	Public with specific conditions	14	Daily	1	CSV
Local Data Bank	Public with specific conditions	Not specified	Quarterly	Not specified	HTML
Register of plant protection products	Public with specific conditions	16	Annual	2035	XLS

Information on the number of applications submitted in CEIDG	Public with specific conditions	7	Monthly	30	CSV, XLS, ZIP

2. Data access methods

There are three main methods of accessing to data sets available on open data portals, not only from Socrata, but also on other portals [14]:

- Access to data by the API,
- Export of the data set (csv, html, xml, xls, rdf, etc. formats),
- Reading data by the html browser.

2.1 Access by API

One of the ways to read data from open data portals is the API - Application Programming Interface, which allows communication of our written application with open data servers. Depending on the chosen portal, various API creation systems are used - CKAN [16] for the EU Open Data Portal or SODA for Socrata [14]. The format of queries and answers is JSON. API is very often used in network applications, where, for example, the http query is written on the server address, and the corresponding value is returned in response. Fig. 6 shows part of the settings of the sample API written in the SODA system [14].

Fig. 6. Part of the settings of the sample API written in the SODA system [14].

Very often when writing API for specific data sets, only interesting data are selected, e.g. only records from the time interval or containing a sequence of characters. For SODA API queries, for example can search by column name:

```
https://data.consumerfinance.gov/resource/jhzv-
w97w.json?product=Tekst do znalezienia
```

We must include query limits when creating the API. Most portals set a quota limit for their data sets that can be used within an hour. For users who do not have a token, the maximum number of queries usually does not exceed 100 queries per hour. API has two methods of accessing to its data: authorization and token. Authorization allows for data management, e.g. adding, deleting, and modifying data sets. The token allows the application to read data from data sets. Access to data sets using the Token in the SODA system is carried out by function shown below [14]:

```
https://data.consumerfinance.gov/resource/jhzv-
w97w.json$$app_token=APP_TOKEN
```

2.2 Database export to available formats

The second way to work on data sets is export of databases to available formats, or to download and analyze databases using program frameworks [3, 14]:

- RS for the R language;
- SoDA-php, Ckan_client for PHP language;
- Socrata, CKAN for the Ruby language;
- SODA, Ckan_client for Java;
- SODA2, ckanjs for JavaScript.

Depending on the portal e.g. Socrata or EU Open Data Portal, the data set provider determines which data format will be made available to the user. These can be various data formats such as: JSON, CSV, HTML, XML, RDF, MS-ACCESS, MS-EXCEL, TXT, OCTET-STREAM or ZIP. Most of applications, e.g. R, can work on JSON, XML or CSV formats but cannot handle the OCTET-STREAM format. If one is using Python for the SoDA format, the following functions should be used [14]:

```
import pandas as pd
from sodapy import Socrata
client = Socrata("domena_zestawu-danych", None)
results = client.get("identyfikator_zestawu-danych",
limit=99999)
results_df = pd.DataFrame.from_records(result_list)
```

For the .NET platform, the data set in SoDA format can be read using [14]:

```
using System;
using System.Linq;
using SODA;
var client = new SodaClient("domena_zestawu-danych ",
"TOKEN");
var dataset = client.GetResource("identyfikator_zestawu-
danych ");
var rows = dataset.GetRows(limit: 99999);
Console.WriteLine("Dumping first results:",
rows.Count());
foreach (var keyValue in rows.First())
{    Console.WriteLine(keyValue);  }
```

Depending on the data format being viewed, the same data is presented in a different ways. When working on a CSV data set, the first line contains the column names (Fig. 7a), whereas in the JSON file each column is saved in value lists (Fig. 7b).

a)

```
Registration Number,Registration Date,Trademark Description,Correspondent Name,.
102,03/15/1968,"""SR""  MONOGRAM",SUNRIVER RESORT LIMITED PARTNERSHIP,11777 SAN
103,03/15/1968,"""SUNRIVER""",SUNRIVER RESORT LIMITED PARTNERSHIP,11777 SAN VIC
158,01/24/1969,THREE TENNIS BALLS AND FOUR TENNIS RACQUETS,"WEST HILLS RACQUET
",C/O NORTH PACIFIC MANAGEMENT INC,1905 SE 10TH AVE,PORTLAND,OR,97214,http://re
272,10/26/1970,G.I.  TYPE CHARACTER  OF A  MAN WITH A JARHEAD - TAKE OFF OF A M
3606,02/01/1935,"""PYREX""",MICHELE N KEEFER-MEHLENBACHER,CORNING INCORPORATED,
3957,04/21/1936,"""CORNING""",MICHELE N KEEFER-MEHLENBACHER,CORNING INCORPORATE
4020,08/29/1936,"""MOLY-KROME""",PACIFIC MACHINERY & TOOL STEEL CO,3445 NW LUZO
4622,10/06/2000,"THE WORDS ""CAPPUCCINO COWBOY"" WITH A LOGO OF A BUCKING HORSE
```

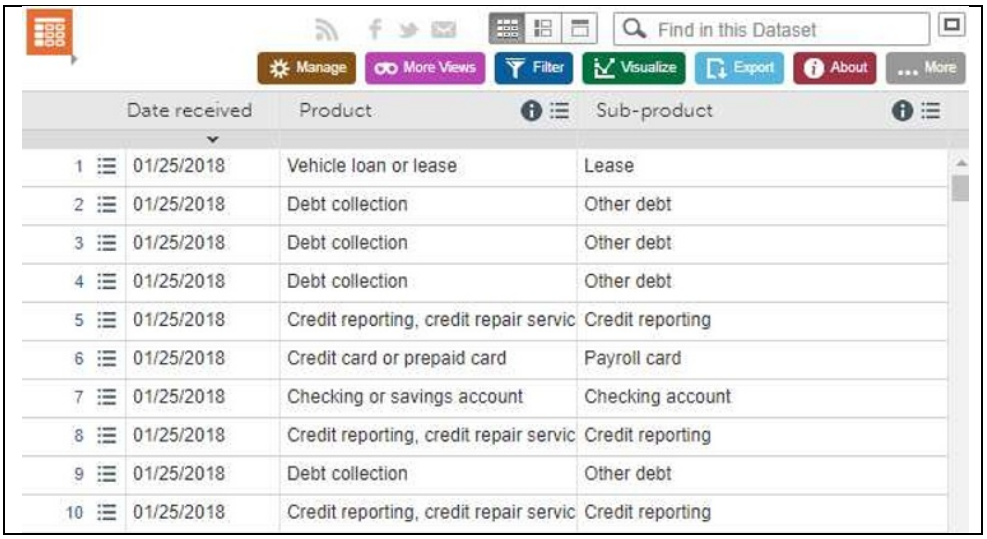
b)


```
"meta" : {
  "view" : {
    "id" : "ny3n-dx3v",
    "name" : "Active Trademark Registrations",
    "averageRating" : 0,
    "category" : "Business",
    "createdAt" : 1296662567,
    "description" : "(Updated 1/2/2018) (Must use
    "displayType" : "table",
    "downloadCount" : 6993,
    "hideFromCatalog" : false,
    "hideFromDataJson" : false,
```


Fig. 7. Data saved in CSV (a) and JSON (b) format [14].

2.3 Data access by the html browser

The third, and the least used method of reading and managing data is accessing data through a browser offering the so-called human readable format. It is simply the ability to preview the whole set of data in the html browser without downloading it. For each set of data there is a description and type of each database column with example values. This method is used only to check whether ones are interested in the data contained in this set, or want to look at what type of columns/records is interested in to copy data to the API. Fig. 8 shows a sample data available throughout the html browser.



	Date received	Product	Sub-product
1	01/25/2018	Vehicle loan or lease	Lease
2	01/25/2018	Debt collection	Other debt
3	01/25/2018	Debt collection	Other debt
4	01/25/2018	Debt collection	Other debt
5	01/25/2018	Credit reporting, credit repair servic	Credit reporting
6	01/25/2018	Credit card or prepaid card	Payroll card
7	01/25/2018	Checking or savings account	Checking account
8	01/25/2018	Credit reporting, credit repair servic	Credit reporting
9	01/25/2018	Debt collection	Other debt
10	01/25/2018	Credit reporting, credit repair servic	Credit reporting

Fig. 8. Open data access throughout the html browser [14].

3. Conceptual model of the open data system

There are specific models to represent Open Data Systems (ODS). UML diagrams are used for it. However, a context diagram is of great help. Such a model is shown in Fig. 9 and is based on the analysis of existing ODS and OGS in Rzeszów, Poland.

Models are built to better illustrate existing or future systems but model will never fully correspond to reality. Modeling is designed to highlight certain features and ignore others; and there is no universal answer to the question of what to distinguish and what to omit. A context diagram is characterized by the following features:

- is a set of actors who are interacting with a given system treated as a single process,
- is a tool used to get to know the scope of system operation and presents the designed system as one process with events taking place,
- defines the system area (system boundary - environment),
- includes persons/organizations/systems - communication (external premises),
- determines data from outside - for processing (flows),
- defines the data generated by the system and transmitted to the environment (flows).

The context diagram shows the general connections between the different subsystems and the external terminators. Functional flows between system functions (in the functional architecture) correspond to physical flows between modules in the physical architecture. Based on the analysis of available sources, many important components can be distinguished that allow the development of a conceptual model for a comprehensive ODS/OGS, as shown in Fig. 9.

In general, we may consider the ODS model in many dimensions. The model depicts the knowledge about the basic functional blocks. One of them is its architectural design. In case of Rzeszów city, it is client- server model. The primary function of the architectural model is to split the entire system into subsystems or components.

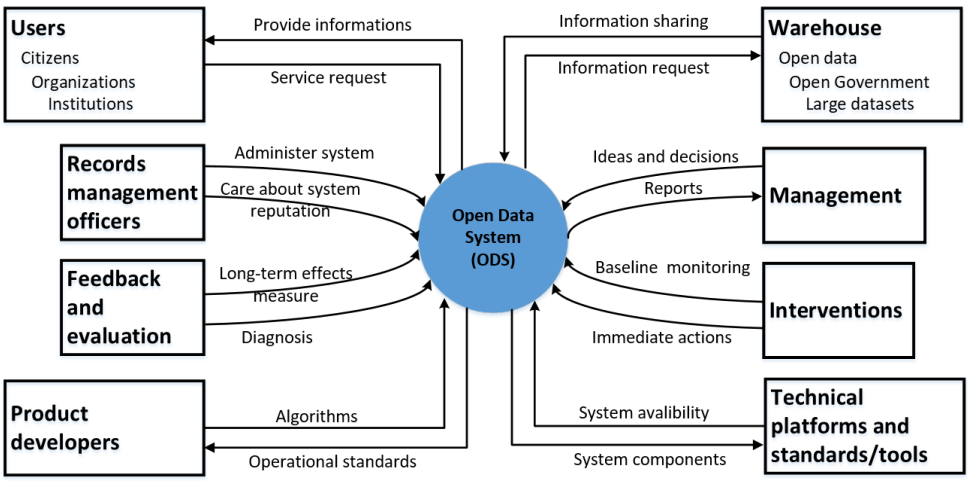


Fig. 9. Main components of the conceptual Open Data/Government System.

The identification of the subsystems and their communication is depicted by Fig. 10. The architectural model for ODS shows transport layer - data is hosted on servers that use HTTP and exchange data throughout the Internet. Clients use the system through an interaction layer that uses the interface and the services generated by the presentation layer. An important part of the system is the access layer because it authenticates to the users and allocates access to the system and thus to the data layer itself, which stores the information resources.

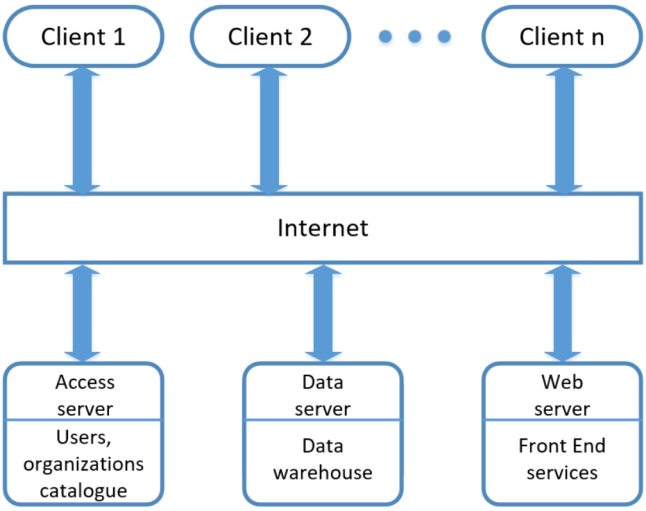


Fig. 10. Architecture of the Open Data/Government System.

The ODS supports business procedures and functions, flow of data, documentation managers, terminators (entities) and databases as well as repositories therefore the repository architecture can be also used. The individual modules of the entire system can be implemented as microservices. It is a proven solution. It enables scalability and flexibility to be achieved without limiting the speed of system development. The main component of the entire system is the database. The data base is a component, which is used to collect, store and process data. Data are usually entered by the staff of individual organizational units. However, some data may be collected directly from the external devices, i.e. sensors using wireless communication solutions. The service is carried out through a client application adapted to various types of user devices - stationary or mobile. The structure of the data warehouse creates further data layers. The lowest layer is made up of data sources. Each subsequent layer is based on the previous one. It is a basic place for storing non-volatile information collected from sources, as well as partial summaries useful in OLAP tasks and decision support. This fits in with the concept of the Big Data solutions discussed previously and the related data processing, advanced data analysis, and decision making support. The data warehouse records the history of the data source, its modifications, and is periodically updated with new condensed information about the current state of the data source, saved alongside previous data sources. Data in data warehouses may be present for the specified time period. The warehouse also performs archival functions.

Another important component of the system is data mart. Data marts are created for the needs of analytical departments, containing selected data in a highly aggregated form, allowing for quick presentation of summaries used in management, long-term planning, historical analyzes, trend analysis, information processing and integrated analyzes. Data warehouses are called thematic wholesalers (data marts, departmental warehouses). Due to the smaller size and the possibility of local work, thematic wholesalers allow for more efficient data handling. They can be implemented as relational databases or special multidimensional structures. Processing of collected data in accordance with analytical algorithms specified by the user (in relation to the source data) is characterized by a delay. It results from the batch process character. The update layer complements the results of batch data processing with data flowing online.

An important part of the entire architecture is the appropriate communication's structure. It is a set of measures enabling the exchange of information between various parts

of the system and external entities. To determine the requirements for a communication system, we have to consider two aspects. Communication channels should ensure adequate speed, bandwidth and cost of transmission. The information will be transmitted without distortion such that it will be properly understood. This aspect concerns the security communication system and transmission protocols. Designing IT systems, we have to fulfill the basic requirements for system and data transfer security. It is necessary to develop a security policy for the system.

Al-Garadi et al. 2018 warns us that "online social networks (OSNs) are structures that help users to interact, exchange, and propagate new ideas [17]. The identification of the influential users in OSNs is a significant process for accelerating the propagation of information that includes marketing applications or hindering the dissemination of unwanted contents, such as viruses, negative online behaviors, and rumors." An important issue of improving the system is obtaining feedback in the Open Data/Government System. The feedback can result in improving its front end, transparency, the speed of work and task handling as well as the possibility of combining data from multiple sources and reusing them. In order to complement internal analytics with information from external sources, system administrators or supervisors may use output data from the open data users as an important source of feedback. It can help them to identify areas in which public services can be improved. This is information of particular importance to the public sector and research institutions. Users can receive more matched offers, data or interface features, as well as a useful analysis of their data by specialized filters and microservices. Data is a resource, and its value and quality are often determined by feedback and validation. Opening data to public may be a way to complement its own analyzes with observations from external sources. A quick and correct response is important in the case of system problems or data security threats. It has a deep sense and influence to product and process improvement. It is also advisable to create an appropriate point of contact for data users or services. It provides assistance and consultation for receiving feedback. It can be used to further improve access to data innovation.

4. Summary

The paper presents current trends in the development of open data portals. Progressive digitalization contributes to more and more effective management of time and resources. An example is the growing number of services available under open government systems. Publishing data in the network are activities aimed at increasing the transparency of the work of offices or institutions. The user can get access to information created by public administration, search and use them in any way. The article aims not only to present ways of accessing open data, the open data system's model or architecture but also to raise the awareness of Internet users about the possibilities that the open data portals provide. The choice of a particular method of accessing data depends on the level of advancement and the needs of a given user. The presented analysis of the danepubliczne.gov.pl or Socrata portals shows that the upward trend is maintained. The number of "open data" is constantly increasing. The increase concerns not only Poland, but also the world. These portals publish credible and confirmed data by the institutions, therefore it is necessary to ensure the confidentiality, accessibility and integrity of such data. It should be remembered that before such data becomes public they must undergo appropriate anonymization processes. It is necessary to adapt the data to ensure the privacy of individuals or institutions. Access to the data requires acceptance of the conditions for the re-use of information, which may differ for each data set.

References

1. J. Gurin, *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*, Published by McGraw-Hill Education, (2014)
2. https://okfn.org/opendata/, (2018)
3. A. Young, S. Verhulst, *The Global Impact of Open Data*, Published by O'Reilly Media, (2016)
4. https://www.europeandataportal.eu/pl/resources/training-companion/open-data-formats, (2018)
5. B. Ubaldi, *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*, OECD Working Papers on Public Governance, No. 22, OECD Publishing, (2013)
6. https://danepubliczne.gov.pl, (2018)
7. https://www.opendatanetwork.com/, (2018)
8. Z. Gomolka, E. Dudek-Dyduch, Y.P. Kondratenko, From Homogeneous Network to Neural Nets with Fractional Derivative, Lecture Notes in Computer Science, vol 10245 (2017)
9. Z.Gomolka, B.Twarog, E. Zeslawska, A. Lewicki, T. Kwater, Using Artificial Neural Networks to Solve the Problem Represented by BOD and DO Indicators, Water 2018, 10(1), 4
10. http://dataportals.org, (2018)
11. P. Colpaert , S. Joye, P. Mechant, E. Mannens, R. Van de Walle, *The 5 stars of open data portals. In Proceedings of the 7th International Conference on Methodologies*, Technologies and Tools Enabling E-Government (MeTTeG13), University of Vigo, Spain (pp. 61-67),
12. Z. Gomolka, B. Twarog, J. Bartman, Improvement of image processing by using homogeneous neural networks with fractional derivatives theorem, Discrete and Continuous Dynamical Systems- Series AI, ssue SUPPL., Pages 505-514, (2011)
13. Z. Gomolka, B. Twarog, E. Zeslawska, Cognitive Investigation on Pilot Attention During Take-Offs and Landings Using Flight Simulator. Lecture Notes in Computer Science, vol 10246. (2017)
14. http://socrata.com, (2018)
15. http://worldbank.org, (2018)
16. http://docs.ckan.org, (2018)
17. M. A. Al-Garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, G. Mujtaba, M. U. Shahid Khan, S. U. Khan. *Analysis of Online Social Network Connections for Identification of Influential Users*. Comput. Surveys 51, (pp. 1–37), (2018)