STATE OF ROMANIAN OPEN DATA IN 2017

Octavian RINCIOG[1], Vlad POSEA[2]

Governmental open data are data published by government and public institutions, under an open license which allows everyone to use this information, without any obligations. In this article, we carry out a study of the existing public data published on Romanian open data portal and we analyze the score of the Global Open Data Index in Romania. We analyze which public institutions publish more data and which are the popular file formats under which the data are published. Our study also focuses on how these open data are used and we present a couple of applications developed using these data. The conclusion of this article is that in order to use this open data to their full potential, they must be transformed in a machine-readable format, such as RDF.

Keywords: Open Data, Global Open Data Index, Resource Description Framework, Government Data

1. Introduction

The definition of Open Data, according to [1] is: "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)". This interpretation is inspired by MIT license [2] used in open source code redistribution. The problem with this definition of open data is that it describes only the openness of data and it does not mention anything about the data quality. Therefore Bonina [3] provides a qualitative definition, specifying that data must be "accessible, intelligible and useable". Boulton [4] adds another property to the data quality: open data should be assessable.

The advantage of this type of information is to encourage transparent and objective assessment of various decisions taken by public institutions. Countries' governments are the largest public institutions that produce and consume data. Thus, in order to fulfill the citizens' desire for a more transparent government, these must make all their data free to access and use.

In Romania, there is a huge need of openness, in terms of data and decisions. In order to receive the citizens' trust and to promote a more transparent

[1] Faculty of Automatic Control and Computer Science, University POLITEHNICA Bucharest, Romania, e-mail: octavian.rinciog@cs.pub.ro

[2] Faculty of Automatic Control and Computer Science, University POLITEHNICA Bucharest, Romania

way of thinking, all public institutions should open their databases. A closed data environment increases the perceived corruption in these public institutions [5][6].

Open Knowledge Foundation (OKFN) developed a methodology, called Global Open Data Index for assessing the state of open data world-wide. This index reflects how data from various domains, such as economy, health, education or legislation are freely available for anyone. According to this index, the score of Romanian Open Data is 51%, being ranked the 24th place out of 122 countries. The leader of this index is Taiwan, which has a score of 90%.

This paper is structured as follows: first we introduce the state of open data world-wide, describing an assessment methodology developed by Open Knowledge Foundation. Also, some of the problems with the current method of data publishing are shown, suggesting that the data should be published as Linked Open Data. Second, we examine the status of Open Data in Romania at the end of 2017 and also the evolution of this paradigm in this country from 2013 until now.

2. State of the art

Most countries developed portals through which governments publish open data. Table 1 presents six such portals and their most important properties, being retrieved in November 2017.

Table 1

Open Data Portals in 6 countries

Region/ Country	Web address	#Datasets	#Institutions	#RDF datasets	#Applications
USA	http://data.gov	197 489	185	8 633	76
UK	http://data.gov.uk	40 893	1 409	138	411
France	http://data.gouv.fr	21 426	1 054	16	1 584
Deutschland	http://govdata.de/	18 618	-	6	22
Japan	http://data.go.jp	17 861	22	0	21
EU	http://publicdata.eu/ dataset.html	47 863	-	2 032	85

What can be concluded from Table 1 is that the number of datasets published as Linked Open Data in all 6 regions is very low, the highest percentage of existing datasets in RDF format being in USA with about 4,4%.

Shadbolt et al. identify [7] the most important benefits, but also some problems that can arise from transforming data into Linked Open Data format. They have identified in this study the following problems:

- ***Data publication*:** There must be various datasets covering as many areas as possible. These datasets should provide a large number of linkage

possibilities between them. Due to data properties, most of these datasets can be linked through common geographical or temporal properties. Being published by different institutions in different formats, these properties can have differences in terms of precision and accuracy.

- ***Data processing*:** In order to comprehend each published dataset, it must be characterized by several metadata expressed in a standard format. Also, data transformation must be done taking into account a common ontology so that all datasets to be expressed using the same vocabulary and not using specific terms for each field, creating islands with no possibility of contact.
- ***Data consumption*:** Viable business model scenarios for implementing and deployment of applications using such data must be created.

On the other hand, the advantages of publishing data in this format (Linked Open Data) consists of: (i) *Transparency*. Decisions of public institutions can be measured more easily and also these institutions can explain their decisions in an easier way. (ii) *Correlation*. Data linkage is natural due to the way of expressing them using unique resource identifiers. Thus, different correlations can be found more easily, for example how was affected the air pollution after an increase in the number of cars registered in a given region. (iii) *Correction*. Published data are exposed in a comprehensible language so, the users of applications can improve these data if they find errors in them.

2.1 Linked Open Data

In 2006, Sir Tim Berners-Lee proposed a scheme[3] that references unique each physical or virtual entity and a very simple way to create links between entities. Using this schema, the major problem of open data - how to identify resources within published data sets - has been solved. In order to accomplish this, he proposed that each entity should have assigned an unique URI and a number of associated properties. Linking data can be carried using these URIs. He also proposed a scheme for classifying data, grading the ease of use. So, using this scheme, data is classified on a scale from 1 star (very hard to reuse) to 5 stars (via URI are easily accessible and have links between them). Fig. 1 presents this classification scheme and specific data formats for each step.

[3] http://5stardata.info/en/

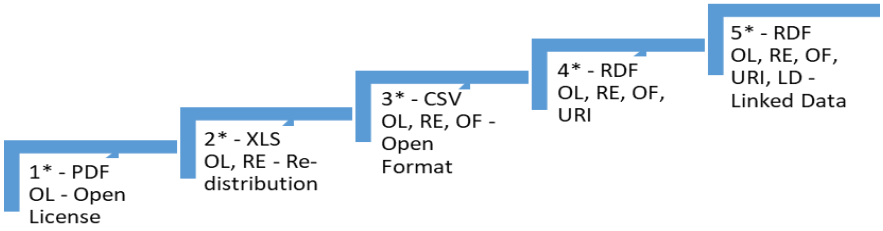

Fig. 1 Five star schema proposed by Sir Tim Berners-Lee

3. OKFN index

Open Knowledge Foundation implemented in 2013 an index[4] for assessing the adoption of open data in several countries. As mentioned above, there are two definitions of open data: one that focuses on the aspect of openness and one that focuses on quality and the usage of published data. OKFN in its own index uses the first definition, focusing only on how truly open these data are. This index quantifies the existence of certain properties of relevant datasets.

The evaluated datasets and quantified properties for each dataset for 2016 are shown in Table 2 and Table 3.

Table 2

Evaluated datasets in OKFN 2016 Index

National Statistics	Election Results	Location datasets	Land Ownership
Government Budget	National Map	Water Quality	Air quality
Government Spending	Pollutant Emissions	Government procurement tenders	
Legislation	Company Register	Weather forecast	

Table 3

Evaluated properties in OKFN 2016 Index

Property	Score	Property	Score
Is data available in machine-readable format?	20	Is data available without registration?	15
Openly licensed?	20	Is data available for free?	15
Available in bulk?	15	Is the data provided on a timely basis?	15

The index for each country is calculated using the formula shown in Equation 1, evaluating datasets for one particular year. Thus, the highest score available is 100% and the lowest score is 0%. At the time of writing this article, obtained indexes are available for years between 2013 and 2016, for 94 countries.

[4] https://index.okfn.org/

For 2016 datasets, the country that has the highest score is Taiwan (90%), followed by the Australia (79%).

$$Index_{country} = 100 * \frac{\sum_{i=1}^{14} score_{dataset_i}}{1400} \quad (1)$$

The scores for this index may vary from year to year for each country due to changes in how is calculated: for example, in 2013, there were only 10 observed datasets with 9 evaluated properties.

Also, another reason for score changing is that the national laws can be changed from year to year: for example, in one year a country can open one dataset, but in the following year this can be denied.

4. Open data in Romania

In Romania the concept of open government data appeared in 2013, when the national open data portal[5] was implemented. Within this portal, the government and various public institutions publish open data that respect an open license, which complies with the definition of open data, i.e. information can be used by anyone for any purpose. In the last two years, there has been an increase in the information published on this site, from a total of 252 datasets in November 2014, to a number of 1076 of datasets in November 2017.

In Romania there is no official regulation stating that the above is the only allowed site where someone can publish open data, the reason for certain public institutions such as the National Statistics Institute publishing on its own website a set of data.

OKFN index is calculated for Romania since 2013 and this index score is presented in Table 4.

Table 4

OKFN Index evolution for Romania

Year	Score	Rank
2013	58%	13
2014	64%	16
2015	58%	15
2016	51%	24

As we can see in this table, the index for Romania had a fluctuating score, with the maximum score obtained in 2014 and the minimum score obtained in 2016. The reason for these changes were:

- OKFN changed over the years how the index is calculated, evaluating more datasets, different properties and more countries.
- More important, Romanian laws have changed during the years regarding the openness of data. For example, in 2015 the statistics

[5] http://data.gov.ro

published by Romanian National Institute of Statistics were truly open for access and usage. But in 2016, these data could be accessed only with a fee and this was penalized by OKFN Index. Another example is that in 2016, Romania did not publish anything about Company Register dataset, which in 2015 was evaluated to 100%.

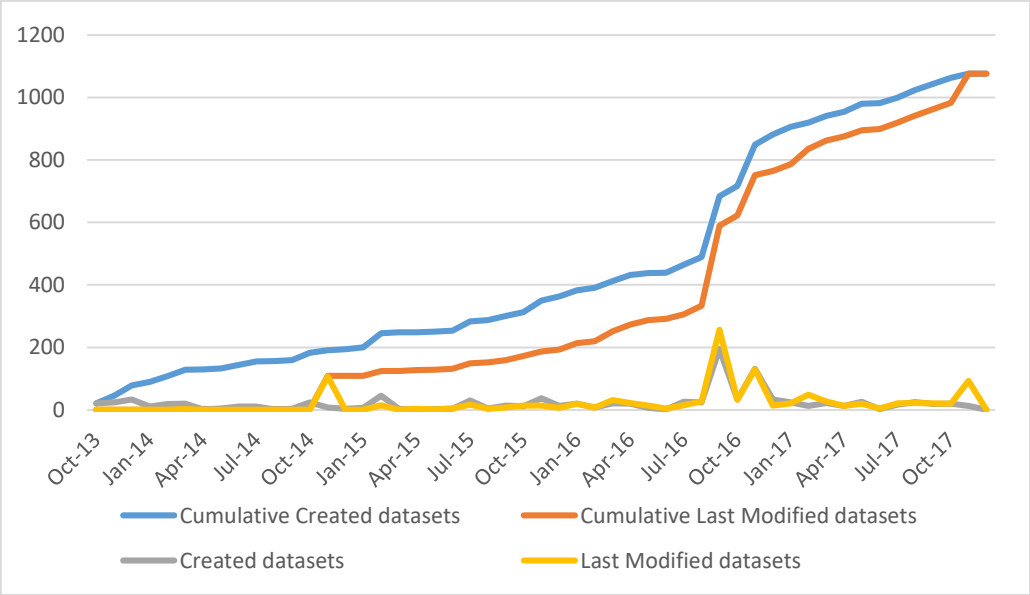

Fig. 2 Total number of published datasets in data.gov.ro

4.1 National Open Data portal

At the time of writing this article, in November 2017, within this portal there were 83 institutions that have published a total of 1076 datasets, comprising a total of 21452 files. Fig. 2 presents the evolution in time of the number of published datasets. As it can be seen from this graph, there have been three different periods for the dataset publication:

- September 2013 - August 2016. In this period the number of datasets grew linearly, with an average of 14 newly created datasets.
- September 2016 - December 2016. The number of datasets grew more rapidly, with an average of 120 newly created datasets
- Year 2017. The number of published datasets return to previous linear growth, with an average of 16 newly created datasets.

To better analyze this information, we created an application that publishes on the blog of the open data laboratory[6] statistics on the number of changed datasets every day. Fig. 3 presents this dynamic, starting from September 2016. As we can see in this picture, there has been an intense activity in months before December 2016, but in the next months the number of average-changed datasets is about 27 datasets.

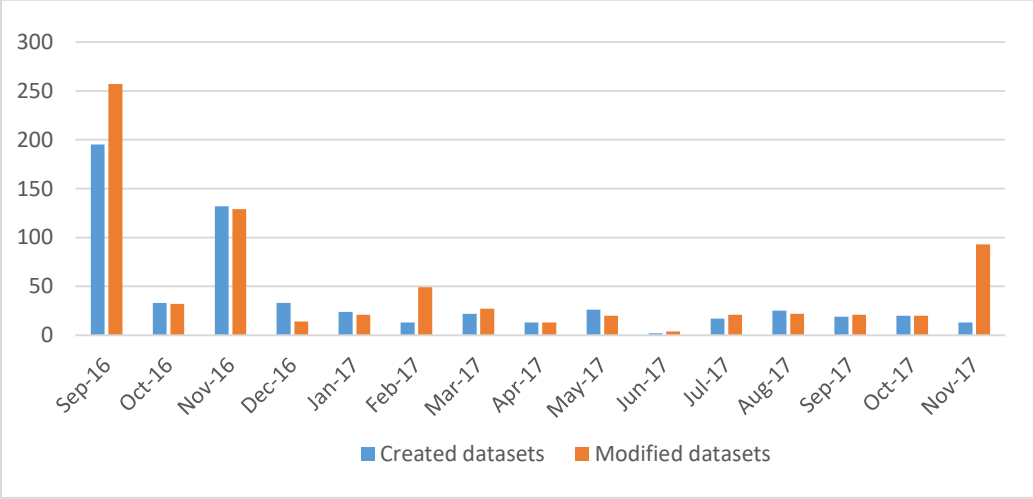

Fig. 3 Number of changes in data.gov.ro between September 2016 and November 2017

4.2 Datasets

The information found in these published datasets can be divided in three distinct categories: (i) *statistical data*, including statistics of several areas such as budget revenues and spending, funds allocated, number of cars in Romania, (ii) *data about physical entities* like museums, schools, pharmacies and (iii) *data about companies*, such as the firms being authorized to perform activities in certain areas. The most important category is the first one, comprising a total of 613 datasets.

As we can see in Fig. 4, the public institution that has published the greatest number of datasets is the Environment Agency, with a total of 216 datasets. The public institutions that have published more than 40 datasets are: Department of Regional Development (211 datasets), Department of Finance (65 datasets), Department of Health (44 datasets), Department of Economy (43 datasets) and Anticorruption Agency (40 datasets). In total there were a number of 52 public institutions that have published data to the national data portal.

[6] http://opendata.cs.pub.ro/blog

The information is found in several file formats, shown in Fig. 5, the most popular being: xls (474), xml (264) and xlsx (225). In total there are a number of 32 file formats.

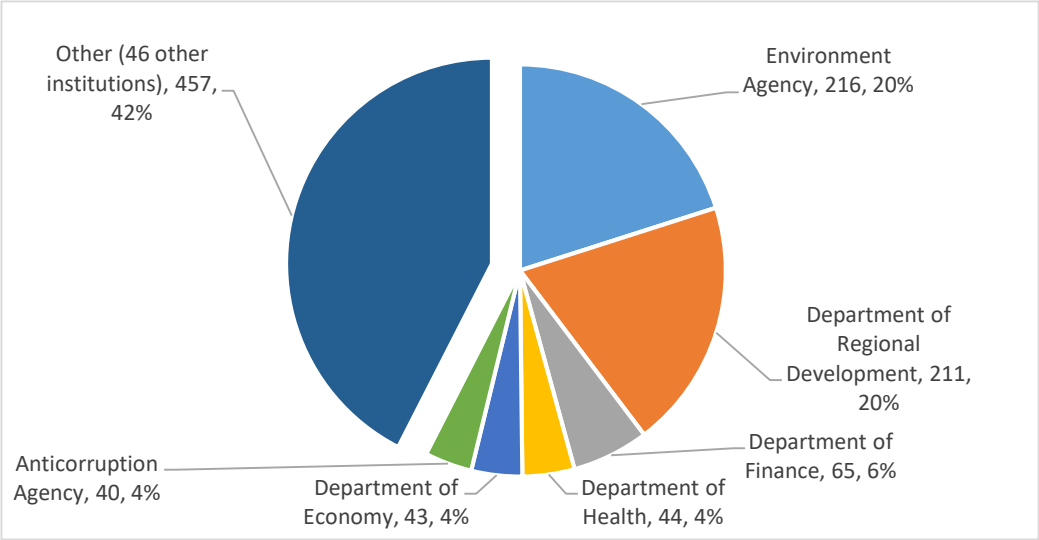

Fig. 4 The public institutions that have published more than 40 datasets

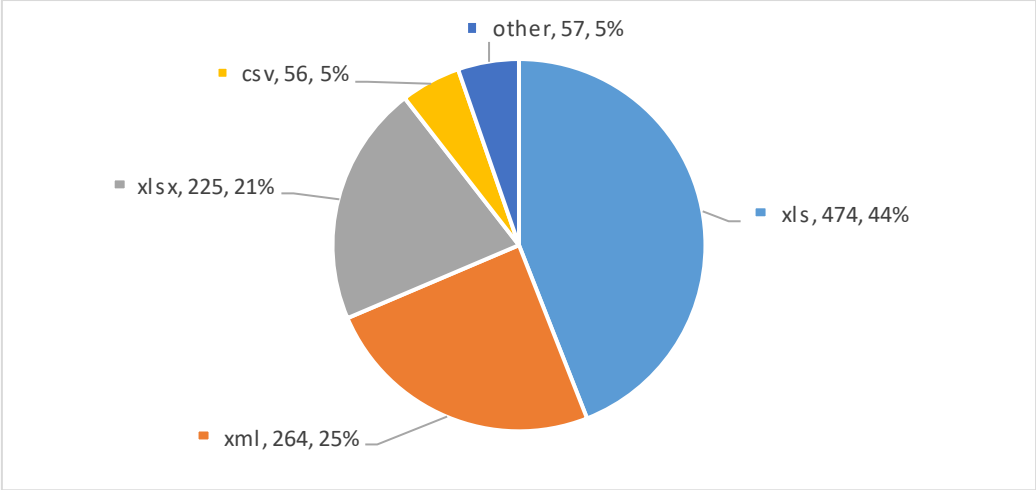

Fig. 5 File formats used for data publishing

4.3 Data quality

The files present on this national open data portal, according to Tim Berners-Lee's classification contain data of one star, two and a maximum of three stars (the most common files formats are: xls, xml, csv). This situation brings two major problems:

- There is no central point of search entities. For example, using current infrastructure there is no easy way to search through all of the results regarding "Bucharest" city, in all published data.
- There is no correlation between data sets. Those who wish to use the data must parse each file and after that implement correlation between various data sources using their own algorithms.

To minimize these disadvantages and using published data on http://data.gov.ro, GovLOD platform [8] was developed. This toolkit transforms data published in several formats in RDF data files, which can be queried using a public SPARQL endpoint. Thus, for the users who are using this platform, the above problems are eliminated.

4.4 Applications

A measure of the usage of public data is represented by the number of implemented applications that are using this information. In Romania, there are two types of implemented applications using open data:

a) Applications that use raw data published by portal or other public institutions. Such an application is Your Money[7], which shows in a graphical way all public Romanian tenders. Through this application, users can find out what procurement contracts concluded a public institution or with what institutions a company has collaborated in time.

b) Applications that use the transformed data in RDF format, following Time Berners-Lee schema presented in section 2.1. Such an application is LODRo [9] application that uses data about physical entities like museums, pharmacies, hospitals and schools converted into RDF. In this application, there were identified 1,014 resources representing Romanian museums, 375 resources representing Romanian hospitals and 15,260 resources representing Romanian pharmacies. These data are linked to the town and country where are physically located. A user of the application can choose the nearest entity from its own current physical location.

5. Conclusions

Open Government Data is vital for a transparent way of thinking for each country. A great amount of open data increases the level of trust of the citizens and offers a predictable method for assessing each decision.

In this paper, we presented a study of existing open data sources in Romania. As we saw, in this country there is a national open data portal, where all public institutions should publish their data. In order to meet people's need, the

[7] http://baniitai.info

number of published datasets must continue the trend seen between September and December 2016.

The changes of the laws regarding the Romanian data openness influences the number of public available datasets. This problem was observed in the OKFN Index, in which Romanian score decreased from year to year, due to denying access to some datasets. Thus, the openness of the datasets is not irreversible, so all the public institutions must continue to publish more and more data in order to see the benefits to all citizens.

We also stated that the actual way of presenting data is not useful for application developers. A more useful way of data publishing is suggested, using RDF triples format. This method helps find all published information about one known entity.

R E F E R E N C E S

[1] OpenDefinition.org, Open Definition, http://opendefinition.org/, Accessed 12 November 2017
[2] Open Source Initiative, The MIT license, 2006
[3] *C.M. Bonina*, "New business models and the value of open data: definitions, challenges and opportunities". RCUK Digital Economy Theme, 2013
[4] *G. Boulton, P. Campbell, B. Collins, P. Elias, W. Hall, G. Laurie, O. O'Neill, M. Rawlins, J. Thornton, P. Vallance and M. Walport*, "Science as an open enterprise". The Royal Society, 2012
[5] *Rajshree, N., & Srivastava, B.,* (2012, July). Open Government Data for Tackling Corruption–A Perspective. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (pp. 21-24).
[6] *Gatti, R.,* (2004). Explaining corruption: are open countries less corrupt?. Journal of International Development, 16(6), 851-861.
[7] *N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, and W. Hall*, "Linked open government data: Lessons from data.gov.uk". IEEE Intelligent Systems, 27(3), pp.16-24, 2012
[8] *O. Rinciog and V. Posea*, "GovLOD: Towards A Linked Open Data Portal", in Proceedings Of The ISWC 2016 Posters & Demonstrations Track. Kobe: CEUR-WS.org, 2017.
[9] *O. Rinciog and V. Posea*, "LODRo: Using cultural Romanian open data to build new learning applications" in The International Scientific Conference eLearning and Software for Education (Vol. 1, p. 267). "Carol I" National Defense University, 2016