OPEN DATA PLATFORMS: ANALYSIS OF THE BRAZILIAN OPEN DATA PORTAL

Shirlei L. O. Carmo, Claudio F. R. Geyer, Julio C. S. dos Anjos,
Institute of Informatics - PPGC/UFGRS
Porto Alegre, Brazil
shirlei,geyer,jcsanjos@inf.ufrgs.br

Abstract

Open data is a concept attributed to the practice of sharing data with anyone, besides being accessed, this data can be manipulated and redistributed. This is a global trend that encourages the transparency of governments and entities in their transactions, as well as providing society with knowledge about relevant data in areas such as infrastructure, health, public spending and the environment. In this work, the authors presented an analysis of the Brazilian Government Open Data platform, presenting metrics about how the organization and quality of the data. The study will show data from the Brazilian platform based on Global Open Data Index (GODI) metrics.
Keywords*: OpenData. Sharing. Data.*

1. Introduction

Transparency, innovation, social and business value addition, participation, impact and engagement of society [1] are justifications for the need for openness, described in Open Knowledge Brazil, a civil organization that promotes free knowledge in various sectors of society.

According to the Open Knowledge Foundation's Open Definition project, the term open in the context of open data and open content means that data can be freely accessed, used, modified and shared by anyone, for any purpose - subject, at the most, to requirements that preserve their origin and openness [2].

Open Knowledge Brasil also defines the main conditions for the data to be considered open. In short, they are: availability and access, that is, they must be available in full and at the cost only of copying, also in a convenient and changeable format; reuse and redistribution, ie in addition to being reusable, it must be possible to combine it with another data set; and universal participation, ie.: everyone should be able to use without any discrimination against persons, groups or fields of action (such as for non-profit or educational purposes only) [3].

Thus, this paper aims to make a sample analysis of the Brazilian Government Open Data platform. Metrics such as amount of resources per dataset will be exposed, if these resources are available for download, the most used data types, if they have exposed licensing in the dataset and the type, also the update periodicity will be checked. These metrics build on the GODI methodology (which is a global reference for open data publishing) and guides institutions that want to open their data.

In the next sections, a contextualization of the use of open data will be presented, briefly exposing the difficulties in the data availability in Brazil, the types of data relevant to society, will be demonstrated how the data were analyzed and the result of the study. At the end, the work will be completed and the possibilities for future work will be verified.

2. Development

In the next subsections will be demonstrated the use and importance of having open data, and exposed the methodology and results of the study.

2.1. Importance of opening data

In addition to the benefits already mentioned, such as transparency of public spending and social engagement due to data exposure, there is also their use in developing solutions, such as *Para onde foi meu dinheiro*, which used data from OpenData-BR, the Brazilian open data portal. The site allows us to see the distribution of government investments in thematic areas such as education, health, social assistance, work, among others [4]. In the open data portal, by the Brazilian data portal application page [1] , you can ac-

1 http://dados.gov.br/aplicativos

cess solutions created for society with open data.

2.2. Difficulties encountered in providing data

According to a report produced by FGV / DAPP in partnership with Open Knowledge Brazil in 2016, the main problems in Brazilian datasets are: incomplete and outdated dataset; unavailability of open format; difficulty working the data; access restriction; difficulty locating data; full base download unavailable; and non-transparent license.

It is possible to check the full analysis of Brazilian datasets through opendatabarometer site [2] , a channel developed by the WordWideWeb Foundation that takes a global measure of how governments are publishing and using open data for accountability, innovation and social impact.

2.3. Data types

Many types of data have potential use and application, such as in the areas of culture, science, finance, statistics, climate, environment, and transport [5]. In VisPublica, Brazil's public data visualization portal, which aims to investigate and apply Information Visualization (InfoVis) techniques to facilitate transparency of public data and decision making [6]. It is possible to check the employment data registered from 2002 to 2010, as shown in Figure 1:

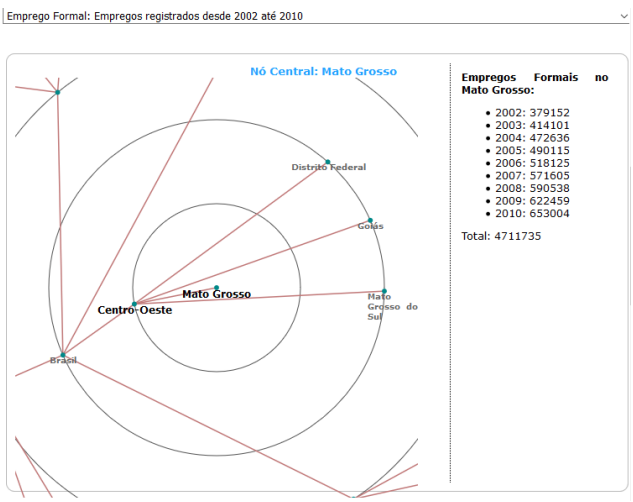

Figure 1. Formal jobs between 2002 and 2010 in the state of Mato Grosso

2 opendatabarometer.org

2.4. Methodology

For data analysis, we used the CKAN API, the largest open source data portal platform in the world. CKAN is a data management system that provides tools for publication, sharing, localization and use of data and which, in this work, was used for data exposure [7].

The technical structure used the GET-able API function and the data was extracted through CKAN's Action API with the following features:

JSON formatted lists of datasets
/api/3/action/package_list , returns available dataset set by platform;
/api/3/action/group_list, returns the groups that are contained in the datasets;
/api/3/action/tag_list, return subdivision of tags by dataset;
Full JSON representation of a dataset, resource, or object
/api/3/action/package_show?id=adur_district_spending, returns the representation of the dataset specified in param 'id'.
Search for packages or resources matching a query
/api/3/action/package_search?facet.field=[%22fieldType%22]&rows=0.

Figure 2. CKAN API features

A processing prototype with the Python programming language was developed through the Jupyter development platform. Data analysis metrics were based on the GODI methodology. This methodology helps to assess data openness in institutions around the globe, and makes an annual progress report on this openness.

According to the methodology mentioned, each data category must contain at least three characteristics. And the characteristics must contain: the required dataset content; data aggregation level; and whether the dataset is up to date.

Only these conditions being met the dataset is analyzed. In this methodology, there is also a questionnaire with 11 questions, 6 with punctuation, which together generate a dataset evaluation score.

Following are the questioner's questions: *Is the data collected by government (or a third-party related or linked to government)?* ; *Is the data available online without the need to register or request access to the data?* ***[15 points]***; *Is the data available online at all?*; *Is the data available free of charge?* ***[15 points]***; *Where did you find the data?*; *How much do you agree with the following statement: It was easy for me to find the data.*; *Is the data downloadable at once?* ***[15 points]***; *Data should be updated every [Time Interval]: Is the data up-to-date?* ***[15 points]*** ; *Is the data openly licensed/in public domain?* ***[20 points]***; *Is the data in open and machine-readable file formats?* ***[20 points]***; *How much human effort is required to use the data. (1 = little to no effort is required, 3 = extensive effort is required).*

2.5. Analysis and Results Obtained

In this work we analyzed the datasets provided by the Banco Central do Brasil (BCB) because, among the organizations contained in the portal, it provides the largest data set (up to the study date 3115 datasets) and also due to the importance of their data to society.

The results to be verified will be: amount of resources per datasets; of these how much are actually downloadable; the formats of these resources; if the dataset has an open data license; if so,what type of license is it.

One thousand BCB datasets were analyzed, sorted by relevance and update data. A JSON formatted list of the dataset has been created, available from CKAN [3] for extraction, then the data was submitted to the code in python and the returned json manipulated. The next paragraphs will display the results of each metric.

Resource is the data itself, made available in a dataset. For example, in BCB's 'Balance Sheets (IFs and Conglomerates)' dataset there are 5 resources which are, among others, the Individualized Balance Sheet in CSV format and the Individualized Financial Consolidated in HTML format.

Due to the amount of datasets and resources analyzed an average of the quantity was made. The result showed that BCB datasets only average 3 resources per dataset.

According to Opendata Commons, a site that centralizes information about licensing and providing open data, open data licensing means freedom of use and fundamental reuse of the data set.

The license type used for the 1,000 data analyzed was Open Data Commons Open Database License (ODbL): applicable to database schemas, information architecture, and data organization. Requires authoring and sharing under the same license. This means that 100% of the verified datasets use the ODbL license.

Each resource in a dataset, must have an option to download the data. The download option, is usually made by a URL, and this URL should make the download available at one time, without any login or without having to pay to download the data.

In the result obtained were analyzed 3999 resources, of which 2974 was possible to download. That is, 74% of the resources were downloadable. This means that more than 50% of resources can be downloaded. Which can result in manageable resources.

According to GODI, the formats provided for each feature must be machine readable, and readable through non-paid software. Thus formats like Json, CSV and HTML are considered acceptable as they meet both requirements.

In the datasets analyzed, the most used formats were HTML and CSV with an average of 25% of resources in this format, followed by JSON and WSDL with 20% utilization. PDF had one occurrence and XML none.

This means that CSV JSON, which is generally manageable, results in 45% of available resources. That is, on average 45% of resources could be easily manipulated.

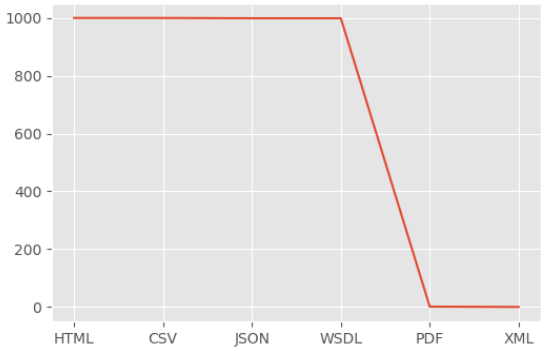

Figure 3. Most commonly used formats in datasets.

3. Conclusions

This work has shown the importance of having data open to society - when it allows transparency of public spending, for example, through resources -, and the use in developing applications - such as vispublica, mentioned in this paper -, or the like, that can generate value and engage citizens.

We evaluated a sample of datasets from a federal government organization, the Banco Central do Brasil, using metrics developed by leading open data evaluators on the globe. And it was found that all data sets analyzed have an ODbL license, over 50% of the features are downloadable and these resources formats,are mainly CSV and HTML.

In the future the data evaluation prototype may be evolved and also the evaluation metrics may be expanded to more control agencies.

References

[1] Open Definition(OD).Definicao de Aberto[online]. Available:http://opendefinition.org/od/2.1/en/. Accessed: 2019-04-20.

[2] Open Knowledge Brasil. Dados Abertos[online]. Available:https://br.okfn.org/dados-abertos/. Accessed: 2019-07-15.

[3] Open Data Barometer. Global Report[online]. Available:https://opendatabarometer.org/4thedition/report/. Accessed: 2019-07-15.

3 api/3/action/packagesearch?q=datasetid

[4] Portal Brasileiro de Dados Abertos. Aplicativos e servicos que utilizam dados abertos.[online]. Available:https://dados.gov.br/pagina/aplicativos. Accessed: 2019-03-15.

[5] OKFN. What is open?.[online]. Available:https://okfn.org/opendata/. Accessed: 2019-07-05.

[6] VISPUBLICA. Sobre o VisPublica. Braslia[online]. Available:https://vispublica.gov.br/vispublica/publico/contato.jsp. Accessed: 2019-05-08.

[7] CKAN. Sobre[online]. Available:https://ckan.org/about/. Accessed: 2019-07-10.