AN APPROACH TO VALUE AGGREGATION IN OPEN GOVERNMENTAL DATA IN BRAZIL

João Pedro Albino
UNESP, Bauru-SP, Brazil
Email: jpalbino@fc.unesp.br

Abstract

The issue of open data can be defined as the paradigm wherein data are available for all services and dissemination. However, the open data debate is just beginning. Open government data are just a point on the way on the road to improve a government's management. However, only data availability is not enough to *add value*. To do so, governments could make a set of tools available, developed on open platform, for a better data treatment and visualization by their users. Thus, the objective of this paper is to discuss a proposal of an "*open*" project (open source), which seeks to foster a debate about open data, big data and display of relevant information by users, providing a repository with coding in an open platform, to assist in the data analysis available in the Brazilian Open Data Portal, thus adding value to such open data. The project proposal uses R language and its *ecosystem* of statistical resources.

Keywords: Data Analysis, Open Data, Language R, Big Data, Brazilian Open Data Portal

1. Introduction and purpose

Open data (Auer et al., 2007), is the paradigm that data may be available for all uses and publication requirements, without copyright restrictions, patents or other control mechanisms. The objectives of the *open data movement* are like the goals of other open access initiatives, such as open source, free hardware, open content and open access (Open Data Handbook, 2015).

The philosophy of open data was established a long time ago (Open Definition Project, 2015), but the definition of the term "open data" itself is recent, gaining popularity with the advent of the Internet and the World Wide Web and, notably, with the launching of numerous governmental initiatives for open data, such as "Brazilian Open Data Portal" (Portal Brasileiro de Dados Abertos (www.dados.gov.br), "Electronic Government Brazil" (Governo Eletrônico Brasil (http://www.governoeletronico.gov.br), "US Government's Open Data" (www.data.gov) in the United States and the United Kingdom's open data (data.gov.uk).

Despite such initiatives, the debate on open data has just begun. The best applications of open government nowadays seek to empower citizens, help small businesses, or create value in a positive and constructive way (Manochaan, 2011).

Open governmental data represent only a point on the road to better education and government, however, other tools are needed to solve other real world problems (Ubaldi, 2013). The Brazilian Open Data Portal is the central point for the search and access to public data in Brazil. The Brazilian policy of open data has as fundamental goals the promotion of transparency, the engagement in social participation, the development of new and better governmental services and the increase of public integrity.

As a strategic action of the open data policy, the federal government is supporting states and cities in implementing local open data policies. Similarly, the Ministry of Planning is convening all federal agencies in public services to publish data and information on the Internet and catalog them in the open data portal.

However, only data availability is not enough to add value. For this, governments could offer a set of tools, developed on an open platform, for a better treatment and visualization by their users. The intention would be to provide flexible algorithms for data analysis, reproducible for some domains based on open governmental data and other relevant data sources. One of the ways for the implementation of this open platform initiative could be the application of the "R" programming language. R programming language is an important tool for the development of analysis and machine learning in numerical spaces. With computers generating more and more data, the popularity of the "R" environment has grown exponentially (Krill, 2015).

In addition, because it is a flexible and open environment, the R language includes in its ecosystem many codes ready for a wide variety of statistical techniques. The R language environment is scalable and offers a rich functionality for the developers to build their own tools and methods of data analysis (Diakopoulos and Cass, 2016).

Thus, the objective of this exploratory work, even though initial, is to create an "open" project (open source), providing a repository with the codes in free platform, to aid in analysis and display of data available on the Brazilian Open Data Portal, adding value to such data.
Another long-term goal is to build a community of developers interested in big data, data science, analysis and display of open governmental data and related topics, using the lessons learned from similar initiatives in other fields of knowledge. To validate the theoretical concept of the research, this article carries out a proof of concept with open data from a Brazilian federal government organization.

2. Theoretical Framework

Many public organizations collect and produce a wide range of different types of data in order to perform their tasks (Ubaldi, 2013). The extraordinary quantity and centrality of the data collected by governments make such data particularly significant as a resource for increasing public transparency. Open Governmental Data (OGD) can be used to help the public better understand what the government does and how well it executes its strategies and make it responsible for illegalities or unfulfilled goals. This is especially true since a considerable amount of this governmental data is becoming progressively more accessible and can be aggregated to information from other sources, such as private information.

In addition to increasing government transparency and sensitizing the general public to government programs and activities, the opening of data can also help generate insights on how to improve government performance (OECD, 2012). Data transparency has increased the basis for public participation and collaboration in the creation of innovative value-added services. Moreover, with the opening of the data, the decision-making of both governments and citizens is, eventually, expected to be improved. Citizens expect to be able to use government data to make the best decision and improve the quality of their lives (Ubaldi, 2013). The government can make available specific databases easily accessible through mobile applications by informing them of their options.

On the other hand, governments are expected to be able to easily access a wide range of data to promote decision-making based on particularities. Finally, OGD are also seen as an important source of economic growth, new forms of entrepreneurship and social innovation (Ubaldi, 2013).

Data are defined and elaborated to transform public service (Manochaan, 2011). The use of open data may help inflict political responsibility, give citizens accountability, and even save lives. It is a commendable and important plan, which is certainly a step in the right direction for governments, the author infers. One of the most significant aspects of "*Open Data*" is the government's definition of what are "free" and "open" data (Manochaan, 2011).

However, there is a risk of depreciating this information because its real and potential value is not appreciated. Data is an asset and the data held by the government make it an extremely valuable asset. Thus, the value of these data begins to reveal itself when different relevant data sources are combined.

Therefore, transparency in OGD is an important feature, but the real value of data can only be measured if properly exploited because they need to be analyzed in order to become useful, and this should be the next step for governments on their journey towards openness and better provision of public services. When data begins to be analyzed and used for innovation, optimization, forecast and prediction, things begin to evolve in regard to information transparency (Manochaan, 2011).

A proposal for open data treatment is the use of development platforms language also created in open source. The R programming language has been considered an important tool for machine learning and analysis. The "R" language has an ecosystem with codes ready for various statistical techniques, as well as flexible and scalable (Diakopoulos and Cass, 2016).

3. Why use R?

In the market today, there are many popular statistical packages and graphics available such as Microsoft Excel, SAS, IBM SPSS, Stata and Minitab (Diakopoulos and Cass, 2016). What reason would there be for this project to use the R environment? Kabacoff (2015) identified many characteristics in the R language that recommend its use:

- Most commercial statistical software platforms are expensive. R is an open source package, which makes it extremely suitable for governments, scientists and users in general.
- R is a comprehensive statistical platform offering all types of analytical data techniques. Any kind of data analysis can be done in R.
- R contains advanced statistical routines not yet available in other packages. In fact, new methods become available for download on a weekly basis.
- R offers state-of-the-art graphics capabilities. For complex data display, R has a comprehensive and important feature set available.
- R functionality can be integrated into applications written in other languages, including C ++, Java, Python, PHP, Pentaho, SAS and SPSS.
- R can run on a wide variety of platforms, including Windows, Unix, Linux and Mac OS X.

However, the R language has pros and cons that its users need to know (Krill (2015). The language has positive points, such as:

- It is very easy to program on more than one level of computer science;
- It has become faster over the years;
- It has served as a language that can bring together different sets of data, tools, or software packages; and
- The possibility to create, analyze and distribute high quality and reproducible code. Because it is a language structured as building blocks, most R programs are collections of scripts organized into projects.

However, despite all its benefits, R has some shortcomings. Memory management, speed, and efficiency are probably its biggest challenges (Krill, 2015). In addition, people coming from other programming environments may also consider the language *peculiar*.

Given the language design, derived from the "S" language from the 1970s, it can sometimes cause problems when used with large datasets (Krill, 2015). Other inherited characteristic is that data needs to be stored in physical memory. However, as computers are being built with an increasing amount of memory, this issue in the future should not be a problem anymore (Krill, 2015).

4. Data Science Process

Data Science (Dhar, 2013), is an interdisciplinary field of processes and systems used to extract knowledge or “insights” from data in structured or unstructured forms. In addition, data science can be considered a continuation of some of the fields of data analysis, such as statistics, machine learning, data mining and predictive analysis, and it is considered similar to Knowledge Discovery in Databases (Leek, 2013).

Data science affects the academy and applied research in many domains such as machine translation, voice recognition, robotics, search engines, digital economics, as well as it has consequences on life sciences, medical informatics, health care Social sciences and the humanities (Leek, 2013). It also strongly influences economy, business and finances. From a business point of view, data science is a part of competitive intelligence, an emerging field encompassing many activities, such as data analysis and data mining (Dhar, 2013). From the data science viewpoint, data represent the traces of real-world processes, and, exactly what traits must be collected, is decided by our data collection or sampling method. After data collection and verification, a simplified and concise mathematical model is used to represent the entire population. This process of moving from the real world to data, and then from the data back to the real world, is the field of statistical inference.

The basic model of this data science process used in this paper will be the model defined in Schutt and O'Neil (2014) and represented in Figure 1.

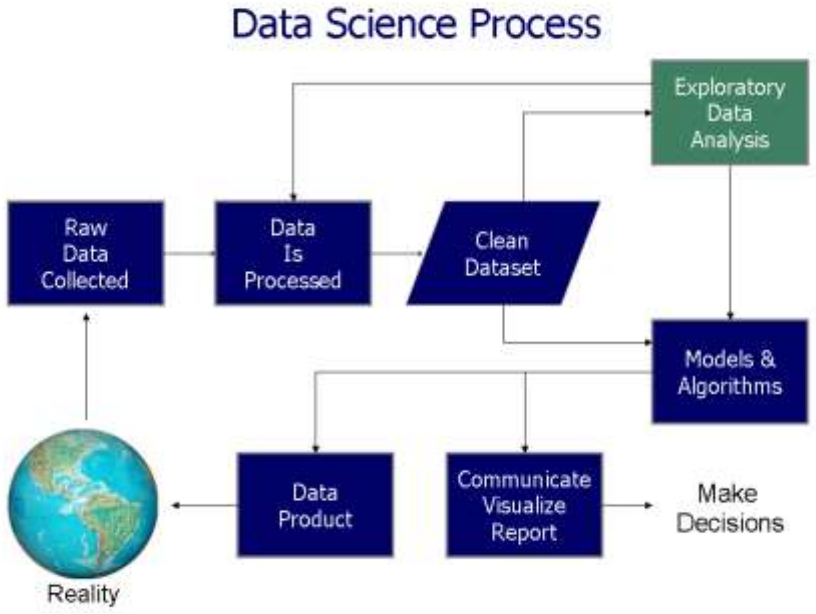


Figure 1: Data Science Process. Source: Schutt and O'Neil (2014).

In short, this model is based on the following steps:

1. Statistical analyzes;
2. Data change to another format (transformation) to be used or processed (analysis, scraping and data formatting);
3. Display (graphs, summaries, tools, etc.).

Following the flow of Figure 1, from Schutt and O'Neil (2014), the data are collected from the real world (*raw data*) and initially must be transformed and *cleaned* for analysis (*data munging*). Once the dataset is cleaned and prepared, Exploratory Data Analysis (EDA) techniques are used for further analysis. Machine learning algorithms are used and, finally, information is represented using data display techniques and tools.

3. Methodology

A research is exploratory when it seeks to establish criteria, methods and techniques and aims to provide information about its object and guide the formulation of hypotheses (Gil, 2008).

Exploratory technological research facilitates obtainment of national and international patents, the generation of wealth and the reduction of technological dependence (Barquette & Chaoubah, 2007). New products and processes can be originated by creative impulses, which from exploratory experiments produce inventions or innovations. The main scientific discoveries are concentrated in exploratory activities, many originated by chance when the observation of a phenomenon occurred during experiments in laboratories.

In this sense, this research is configured as exploratory because it aims, through an "open source" project, create a method and technique for a repository with platform code free, to support the analysis of the data available in the Brazilian Portal of Open Data and thus, adds value to such data. To do so, this project uses the R language and its package ecosystem.
The main research question of this work is:

> "Is it possible to add value into open data through open code platform, supporting its analysis and visualization and building a community interested in analyzing such data, using lessons learned from similar initiatives in other fields?"

To answer this question, the bibliographical method will be used and, likewise, site search in open data sites and in the Brazilian Portal of Open Data, to establish the initial sources of data. For the development of open data analysis and display tools, initial research will be carried out on the ROpenGov and Data.Gov portals in addition to the Bioconductor and rOpenSci sites, among others, to use the experience and documentation of lessons learned.

In regards to the R programming language, the site to be used is The R Project (2015). CRAN-R is a network of FTP and web servers around the world that stores code versions and documentation for identical and up-to-date R's. To minimize the load of the network, CRAN project mirrors located in Brazil will be used at the State Universities of Santa Cruz, Federal University of Paraná, University of São Paulo (USP) and Osvaldo Cruz Foundation (The R Project, 2015).

After establishing the previous points, the site "rOpenGovBr" in beta (initial) should be constructed. After the construction, an invitation to the communities of academic and private research will be elaborated and sent, evoking collaboration towards the site.

4. Data Science Process Application

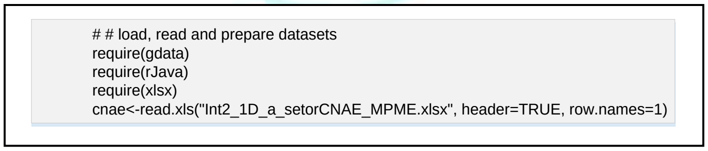
In order to elucidate the concept established in the topics developed in the previous items of this article, we will carry out *a proof of concept* using as a basic model the *data science process* from Schutt & O'Neil (2014) and represented in Figure 1.

According to the Figure 1 framework, the first important step before solving a problem is to define exactly what the (Real World) problem is expected to solve. In the case of a data cognition process, one must be able to translate data issues into something actionable (Schutt & O'Neil, 2014).

In this specific case, we seek to solve a problem for MSEs, trying to determine the possible investment disbursements for the next year in this sector of the Brazilian National Bank for Economic and Social Development (BNDES, 2016), the main instrument of the Brazilian Federal Government for the long-term financing and investment in all segments of the Brazilian economy. Data from previous years (2006 to 2015) are available on the BNDES website (2016)[1]. The main issue will be to predict whether BNDES (2016) will continue to invest in this segment at the same levels of previous years. The important thing at the end of this first step is to get all the information and context needed to solve the problem.

Once the problem is defined, one ought to get the data necessary and build the knowledge needed to transform the problem around a solution. This step of the process involves reasoning which data is needed and finding ways to obtain it, whether it is querying internal databases or acquiring external databases.

In the specific example, the "raw data" in Excel format (.xlsx) was obtained. Figure2 shows the source code snippet in R for the file load in the language environment.
After loading the raw data, it is necessary to evaluate them before any analysis. Often, data can be quite confusing, especially if it has not been collected and stored carefully, containing errors that could corrupt the analysis, such as: values set to null which actually are zero; duplicated values; and *missing values*. It is up to the data scientist to evaluate and verify to make sure the data has accurate information. This step, in Figure 1, corresponds to the processes of transformation (processing) and cleaning of the data.



```
# # load, read and prepare datasets
require(gdata)
require(rJava)
require(xlsx)
cnae<-read.xls("Int2_1D_a_setorCNAE_MPME.xlsx", header=TRUE, row.names=1)
```

Figure 2. R source code to load file.

The next step is to conduct an exploratory data analysis. Exploratory data analysis (EDA) is the first step to be taken for the construction of a model. EDA is a critical part of the data science process and also represents a philosophy or way of performing statistical processes practiced by a breed of statisticians from the Bell Labs Lab tradition (Schutt & O'Neil, 2014). The initial graphs of the exploratory data analysis in the sample are represented in the graphs and histogram of Figure3.

The next step of the exploratory analysis is the *model construction*. In our example, we sought to create a *predictive model* using data already available in the last years of BNDES (2016)

[1] The original worksheet, in PDF format, is available at: <http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/empresa/estatisticas/Int2_1D_a_setorCNAE_MPME.pdf>. Access in: 09/13/2016 17:02:37.

disbursements to small and medium-sized companies from all regions of Brazil and from all business areas.

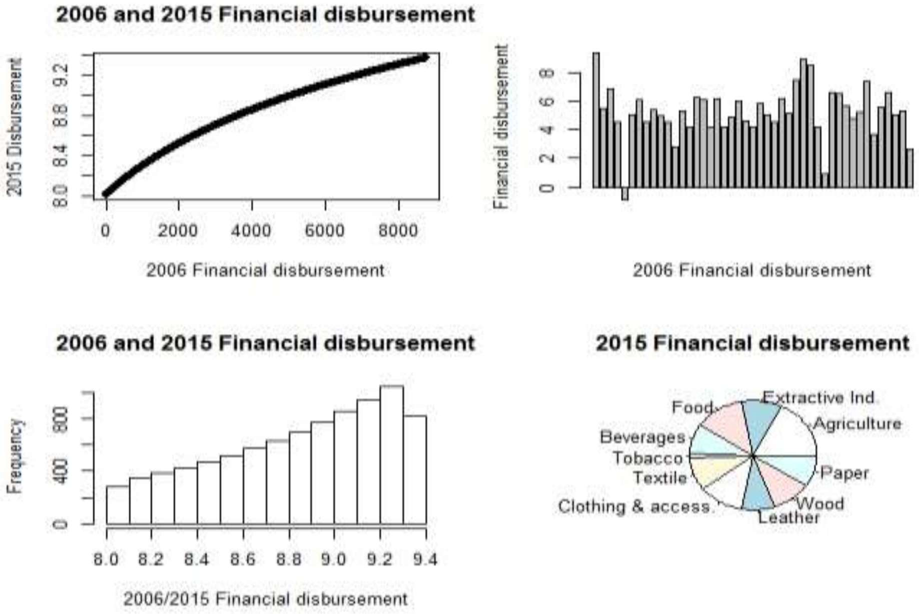
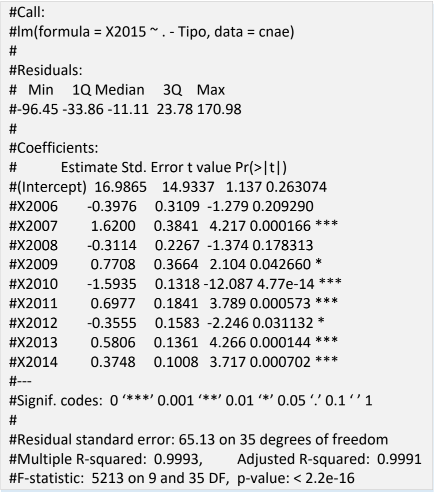


Figure 3: Exploratory Data Analysis

In this step, a regression model was constructed to try to predict disbursement values based on other value attributes (previous years), using the multiple linear regression model.

The linear regression modeling approach usually consists of a response variable, which we try to predict and several input variables. The model also assumes that there is a linear relation between the input variables and our response variable.

In Figure, 4 one can see the details of the model using the *summary* command of the R language in the *initial_model* variable, which gives detailed information about the model, the coefficients of the multiple variables in different metrics.

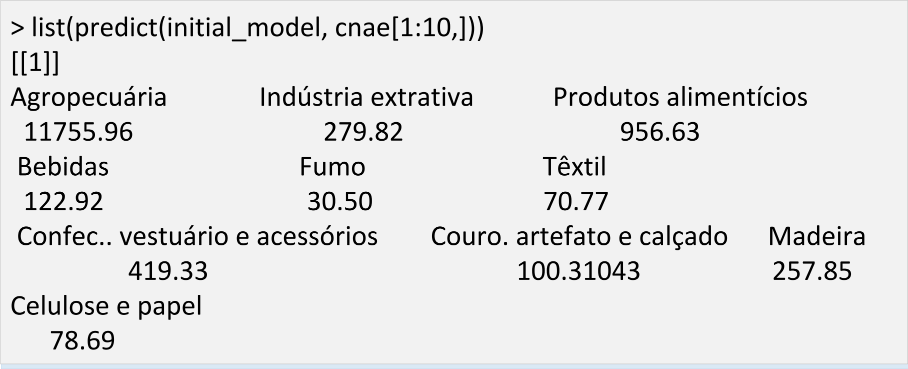


```
#Call:
#lm(formula = X2015 ~ . - Tipo, data = cnae)
#
#Residuals:
#   Min     1Q Median     3Q    Max
#-96.45 -33.86 -11.11  23.78 170.98
#
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)  16.9865    14.9337   1.137 0.263074
#X2006       -0.3976     0.3109  -1.279 0.209290
#X2007        1.6200     0.3841   4.217 0.000166 ***
#X2008       -0.3114     0.2267  -1.374 0.178313
#X2009        0.7708     0.3664   2.104 0.042660 *
#X2010       -1.5935     0.1318 -12.087 4.77e-14 ***
#X2011        0.6977     0.1841   3.789 0.000573 ***
#X2012       -0.3555     0.1583  -2.246 0.031132 *
#X2013        0.5806     0.1361   4.266 0.000144 ***
#X2014        0.3748     0.1008   3.717 0.000702 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 65.13 on 35 degrees of freedom
#Multiple R-squared:  0.9993,       Adjusted R-squared:  0.9991
#F-statistic:  5213 on 9 and 35 DF,  p-value: < 2.2e-16
```

Figure 4: Model summary.

It is observed that the adjusted R² (*Adjusted R-squared*) value is 0.9991, which indicates that 99.91% of the variation of the response variable (X2015) is explained by the input variables. The higher this value, the better the model because it will explain most of the variability observed in the response variable, which is sought to be predicted.

Finally, Picture 6 shows the first 10 business areas of small and medium-sized enterprises and the expected disbursement amounts by BNDES (2016) for the succeeding year.



```
> list(predict(initial_model, cnae[1:10,]))
[[1]]
Agropecuária          Indústria extrativa        Produtos alimentícios
  11755.96                 279.82                       956.63
 Bebidas                    Fumo                 Têxtil
  122.92                    30.50                70.77
 Confec.. vestuário e acessórios       Couro. artefato e calçado     Madeira
          419.33                              100.31043               257.85
Celulose e papel
    78.69
```

Figure 5: Estimated disbursement prediction.

As it can be seen in the previous items, in this article, a proof of concept was carried out with open data from a federal government agency that supports and finances investments in small and medium enterprises with the purpose of validating the theoretical concepts established in the methodology. Based on the results obtained, it was possible to elaborate forecasts on the possible disbursements of investments for the next year in each sector, an important piece of information for small and medium enterprises that depends on this federal resource for their management.

Conclusions and Expected Results

This project aims at starting the debate in Brazil regarding Big Data and Open Governmental Data (OGD) made available by the Federal Government and the need to facilitate the display and interpretation of such data.

With this research question in perspective, the intent is to introduce the R open source programming environment and its ecosystem in order to ease the use of this environment. In addition, the possibility of creating in the country its own community of collaborators to the project, would make the development of a scientific culture of Open Data treatment and its statistical and modeling tools faster.

Thus, based on the lessons learned from similar open projects, the expected results and possible applications for Public Management would be:

- **Statistical and graphical methods.** Providing computational tools dedicated to the analysis of data sources relevant to social sciences, humanities, computing, digital media and related fields.
- **Documentation.** High quality documentation is critical to the usability of the project. **Scalability.** The development of a set of tools by the community ensures that the applicability of the tools extends beyond individual databases and it is compatible with other available tools. Tools and workflows can be easily adapted to different research questions in different locations. Providing a unified interface to the data sets, making their analysis understandable.

- **Reproducible search.** Reproducible search refers to the idea that the end product of academic research is an article and the entire computational environment used to produce the search results, such as code, data, etc., that can be used to reproduce the results and create a new paper based on published research. In the project, the site may offer tutorials and online transparent documentation showing how to automatically generate the obtained results, providing full details and algorithms on how to access, preprocess, analyze and report data and analyzes. This approach serves as a model for good reproducible computing practices.

- **Open source.** For shared version control, one can use the GitHub environment extensively. **GitHub** is a shared hosting service for projects that use the control in addition to offering the features of a social network and graphics showing how developers work the versions of their repositories. In this environment, the contributions will be openly licensed to ensure that the scientific community owns the software tools needed to conduct the research.

- **Open development.** The goal is to turn users into developers, whether by contributing to packages compatible with the project or with documentation. The project will aim at providing a forum that brings together different groups with common goals, often through collaborative development. In this case the use of "Slack" (www.slack.com) and the open platform for communication of group projects, offers resources to develop training for researchers on computational and statistical methods in social sciences, humanities, digital media and computing.

References

i. Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z., 2007, DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web,Lecture Notes in Computer Science*,SpringerLink.com, Volume 4825/2007, pp. 722-735.

ii. Barquette, S, and Chaoubah, A., 2007. Pesquisa de marketing. São Paulo, Brasil, Saraiva, pp. 127.

iii. BNDES, 2016, Estatísticas operacionais para download. [Online]. Available at: http://www.bndes.gov.br/SiteBNDES/bndes/bndes_pt/Institucional/BNDES_Transparente/Estatisticas_Operacionais/estatisticas_download.html. [Accessed: 8 September 2016].

iv. Diakopoulos, N. and Cass, S., 2016, Interactive: the top programming languages, *IEEE Spectrum*. [Online]. Available at: http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016. [Accessed: 8September 2016].

v. Dhar, V., 2013, Data science and prediction, *Communications of the ACM*, 56 (12): 64. doi:10.1145/2500499.

vi. Gil, A. C., 2008, Como elaborar projetos de pesquisa. São Paulo: Atlas, pp. 176.

vii. Kabacoff, R. I., 2015, R in Action: Data analysis and graphics with R, 2nd. Edition, *Manning Publications Co*., 608 pp.

viii. Krill, P., (2017), Why R? The pros and cons of the R language. *InfoWorld*. [Online]. Available at: http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html. [Accessed: 31 January 2017].

ix. Leek, L., 2013, The key word in "data science "is not data, it is science. *Simply Statistics*. [Online]. Available at:http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/. [Accessed: 5 September 2016].

x. Manochaan, I., 2011, On the road to open data, *IDGConect*. [Online]. Available at: http://www.idgconnect.com/blog-abstract/263/ian-manocha-uk-on-road-open-data. [Accessed: 10 September 2015].

xi. OECD, 2013, *Working Papers on Public Governance*, No. 22, OECD Publishing. [On line]. Available at: http://dx.doi.org/10.1787/5k46bj4f03s7-en. [Accessed: 30 September 2015].

xii. Open Data Handbook, 2015, [Online]. Available at: http://opendatahandbook.org/guide/pt_BR/what-is-open-data/. [Accessed: 15 September 2015].

xiii. Open Definition Project, *Definição de Conhecimento Aberto*, [On-line]:http://opendefinition.org/od/1.1/pt-br/. [Accessed: 10 September 2015].

xiv. Schutt, R. and O'neil, C., 2014, Doing data science: straight talk from the frontline, *O'Reilly Media*, California, USA, 408 pp.

xv. The R Project, 2015, [On line]. Available at: https://www.r-project.org/foundation/, [Accessed: 5 October 2015].

xvi. Ubaldi, B., 2013, Open government data: towards empirical analysis of open government data initiatives, *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/5k46bj4f03s7-en.