11th International Forum on Knowledge Asset Dynamics
Dresden, Germany 15-17 June 2016
Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

Strategy for the Automated Diagnostic of the Openness Degree in Government Data

Ramón Reyes-Carrión*, Elio-Atenógenes Villaseñor, Mario Graff-Guerrero

INFOTEC Centre for Research and Innovation in Information and Communication Technologies Circuito Tecnopolo Sur No 112, Col. Fracc. Tecnopolo Pocitos C.P. 20313, Aguascalientes, Ags. México

* *Corresponding author*

Structured Abstract

Purpose –An information extracting and modelling strategy, based in big data analytics, aimed to evaluate in a (semi-)automated manner, the Mexican government data, analysing the pertinence, quality, usability and organization in the open data of the Mexican government.

Design/methodology/approach – We propose a detailed study of the formats, structure, temporality, uniformity, ease of access, available tools, to determine the openness degree of the data of the Mexican government. Additionally, the results obtained from the study, will be useful for the modelling and evaluation of the information to design and implement the strategies of analysis of the open government data, using big data analytics advanced techniques.

Originality/value – This methodology puts in evidence the feasibility of the categories defined by the research group: "Seminario de Investigación de gobierno abierto y big data". Lastly we diagnose the Mexican government open data using big data analytics tools.

Practical implications – The outcome will be the application of big data analytics tools to open data susceptible to be applied to data from developing countries government. The modelling of the data will be the starting point for the design and implementation of the data analysis strategies to be applied to open government data through the use of big data analytics advanced tools.

Keywords –Government Datasets, Open Data Repository, Big Data Analytics, Data Modelling, Openness Degree

Paper type – Academic Research Paper / Practical Paper

11th International Forum on Knowledge Asset Dynamics
Dresden, Germany 15-17 June 2016
Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

1 Introduction

Open government initiatives have emerged historically based on three pillars, namely: openness, transparency and participation. We take as a definition the one that gives the Government of Canada[1].

Open Data is defined as structured data that is machine-readable, freely shared, used and built on without restrictions[1]. In this decade the Open Government Partnership has become a centralizing organization

> *.. In total, 69 OGP participating countries have made over 2,250 commitments to make their governments more open and accountable*[2].

In particular the Mexican government established five objectives with specific goals and actions, to promote transparency and accountability, these include topics such as:

- Social and public security
- Government budget
- Education
- Public infrastructure
- Natural resources

The actions included:

1. The conformation of norms and vigilance committees
2. The creation of web sites and search engines to facilitate the access to the data.

As a local representation of the OGP, in Mexico, the Alianza gobierno abierto (http://gobabiertomx.org/) has references to the commitments and norms of the Mexican government at all the levels.

Officially there is a centralized index at datos.gob.mx, where we based our analysis; however there is a good amount of information not included there, other sources are:

- Detained people
- Budget transparency (partially indexed in "Datos")
- Direct federal spending

There are various international organizations that track, classify and analyse government open data, two of them are:

[1] *(http://open.canada.ca/en/open-data-principles)*
[2] *See more at: http://www.opengovpartnership.org/about#sthash.7HLYTBK5.dpuf*

Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

- Global Open Data Index Based on the preparation, implementation and impact.
- Open Data Barometer Based on level of aggregation and the specificity of the indicators.

They group the data to rank the countries and publish their findings about once a year.

The Mexican government publishes a dashboard with the progress in the implementation of the 26 commitments in the 2013-2015 action plan (http://imco.org.mx/wp-content/uploads/2014/01/pa_aga_2015.pdf).

2 Methodology

The next diagram describes the strategy that we propose to evaluate the feasibility of the application of big data analytics techniques considering the Mexican federal portal of the open government data framework as data source.

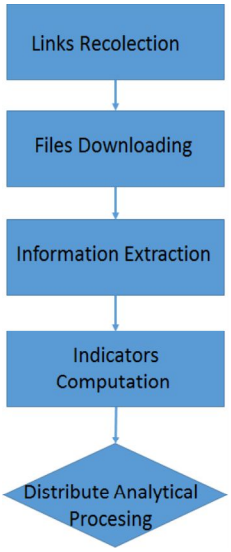


Figure 1. Diagram of the proposed strategy.

The application of this strategy allow us to give a diagnostic of the feasibility and convenience of application of big data analytics techniques as tools for knowledge-based value drivers.

Considering the fact that there are various indices (Alonso, J.M. 2012, Bertot et al. 2012) based on the indicators and the accomplishment of the commitments made; we take

Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

a different approach and focus ourselves on the technical difficulties found when semi-automated tools are used to access and analyse the data. With this in mind we proceed to the evaluation of the following aspects:

- Ease of access: this term refers to the barriers found when accessing and analysing the data itself, and for that we focus in what is presented as a first instance to the public, and the immediate difficulties encountered.
  - Broken links, or links pointing to web pages that don't correspond to the advertised data.
  - Data in proprietary format or that require commercial software to be able to make use of it
  - The codification of the data is not standard
  - Interoperability of the sites so that access is denied, given the policies of the referred system
  - The format of the data is not explicit or does not correspond to the one announced
  - The organization and categorization of the data within the web portal, and the difficulties to find the data.
- Structure of the data: once the data can be accessed and downloaded, there still may be problems to be able to use it, in particular to be subject of analysis with common big data tools.
  - Headers with a clear meaning
  - Uniform headers for each data source (at least)
  - Explicit and consistent units

There are other features that make feasible and convenient, the application of big data analytics techniques, some of those are: level of aggregation, historical data, coverage, treatment of private and personal information; and still need to be incorporated.

3 Diagnostic of Open Big Data from Mexican Government

3.1 Information Retrieval of the Open data catalogue

The primary issue that should to be solved if pretending to offer the service of open data access is the possibility to find the data set that obeys the need for information.

In this work we concentrate on the official index provided by the Mexican government (datos.gob.mx), having in mind that one of commitments is the centralization

11th International Forum on Knowledge Asset Dynamics
Dresden, Germany 15-17 June 2016
Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

of all the digital government initiatives, including the access to the open government data, in a single portal (www.gob.mx).

As a stayed in the presidential "**DECRETO por el que se establece la regulación en materia de Datos Abiertos (DECREE to establish the regulation in the matter of -governmental- Open Data)**"; all Open Government Data should be accessible from datos.gob.mx. For our analysis, we filter, using the site tool, only documents with formats: XLS, and CSV; leaving out the geo-referenced formats: KML, SHP, KMZ. From that, failing to use common crawling techniques, we obtained manually 653 url's that point to the data, these were the input for the scripts and big data analysis tools. In this way we obtained the next figure.

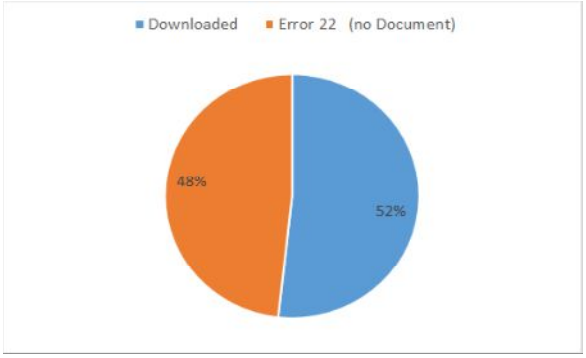


Figure 1. Proportion of downloaded files from the links recollected.

This gives us most of the elements of the ease of access group, namely the broken links, access denied, or nonstandard codification. One possible reason for broken links is the called: electoral ban ("beda electoral"); this is a federal government policy that is applied to all the mechanisms of diffusion and publicity of governmental achievements during federal (and local) electoral campaigns. From the documents downloaded, we have the following distribution of formats.

Considering the fact that we filtered only CSV and XLS, formats. We proceed using only the first group, as far as this is the one properly categorized and that does not require the use of commercial software. Then we move to the structure aspects using the CSV downloaded files, and find the following histogram of the name of the variables:

11th International Forum on Knowledge Asset Dynamics
Dresden, Germany 15-17 June 2016
Published in Proceedings IFKAD2016
ISBN: 978-88-96687-09-3
ISSN: 2280-787X

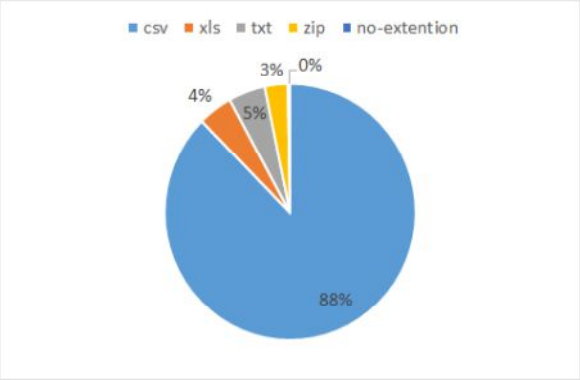
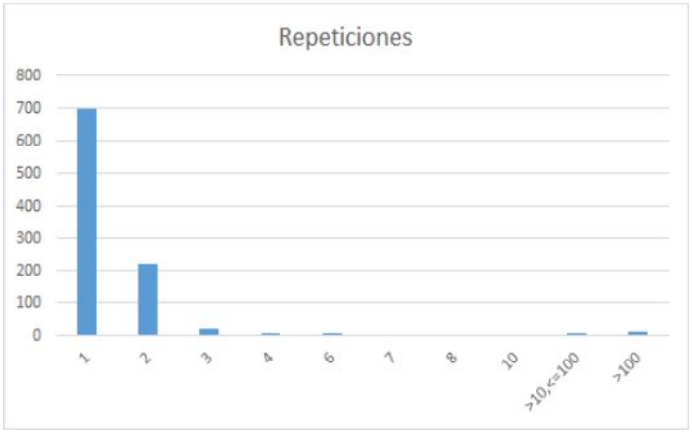


Figure 2. Distribution of format from the dowloaded files.

The Frequency in the number of Occurrences in the names of the variables (for example: "Periodo" (period) 14, "Entidad_Federativa" (federal agency) 12, "Fecha" (date) 11 and "Estado" (state) 10); is an indicator of the possibility of generate big matrices (data derived from data) due to the application of cross-products. That's why this is an important indicator of how convenient would the use of big data analytics techniques.



3.2 How big is open government data?

In the next Figure we can appreciated the number of records in the files (in terms of number of lines). As it can be seen only a little portion of files with more than one million of records.