The integration of multi-source heterogeneous data: an open data case study for budgetary execution in Brazil

Emergent Research Forum Papers

José Rodolfo Beluzo
USP - Brazil
jrbeluzo@usp.br

Gisele da Silva Craveiro
USP - Brazil
giselesc@usp.br

Abstract

This paper presents a proposal for data and schemes integration of public budget execution (revenues and expenses) in a Data Warehouse from the data available in portals transparency. The prototype integration proposal aims better information transparency of federal entities to better subsides researches on public policy, political science and also auditing and citizen participation in public policy, with a case study in budget execution in Brazil.

Keywords

Data warehouse, open government data, heterogeneous data.

Introduction

The need for the integration of data has become of crucial importance for improved information management in both the private and public sectors, where the amount of recorded data in different places and that have their own data sources, is a real-life situation (Halevy, A. Y.; et al.).

One of the problems of data integration lies in the heterogeneity of the data that have to be integrated. This heterogeneity can take several forms: data procedures, nomenclature for fields, domain value, DBMSS or types of files when there is no DBMS and the files are available in a directory of Web archives, either in an open file (e.g. CSV) or proprietary formats (e.g. XLS).

When it comes to the concept of government data, these provide a wide range of information that is of interest to the public. The budgetary information is an essential part of this. Through revenue estimates it is possible to make improvements in several cities of a country (Giacomoni, 2012), as well as to give proof of budgetary transparency to the people, which is a key factor in ensuring greater accountability (OECD, 2002).

In Brazil there is a legal historic framework which makes it mandatory to publish public budgetary records in real time in the transparency portals of the world network of computers – the Internet: a) The Brazilian Constitution of 1988 (Brasil, 1988); b) Law of Fiscal Responsibility (Brasil, 2000); c) Capiberibe Law (Brasil, 2009); d) Access to Information Law (Brasil, 2012). The last one rules that citizens have a constitutional right of access to information provided by the Government. In spite of this, there is a gap which must be filled with regard to the information that can be produced through these data – and it is in this area that the heterogeneous data information is framed.

Currently, federal bodies disclose their expenses and taxes in a document called the “Budgetary Guidelines Law”. The expense and taxes are classified in various ways as laid down by the “Technical Budget Manual” (MTO, 2014), so that they can assist in a) formulating programs, b) putting the budgetary execution into effect, c) providing accountability and d) organizing the data in a way that makes it possible to analyze the economic effects of government activities. (Burkhead, 1971). However, there is no model for the publication of these data in transparency portals.
Budgetary execution in Brazil is divided into three layers of federal bodies: 1 Federal Government, 26 State Governors and 5,570 municipalities (the federal bodies have a good deal of autonomy with regard to their budgetary executions). Thus, apart from the fact that it does not have a single standardized form, the method of classifying the budgetary execution varies considerably from one federal body to another and this makes it even harder to achieve a possible integration of the data.

With regard to the guidelines, in the publication of open government data, Brazil is subject to OGD principles defined by a group of activists in 2007. On the basis of these guidelines, in 2013 a study carried out by Craveiro et al, (2013) collected and analyzed more than 300 datasets published by 88 transparency portals and National Audit Offices. They found that most of the portals did not conform to the standards of open government data and also did not meet the necessary legal requirements.

Nonetheless, there are various works which already make use of open government data in Brazil. An example of these works is “Caring for my neighborhood" which uses datasets of the expenditure of the municipality of São Paulo/SP to map the forecasts of expenses, as well as what has effectively been spent in accordance with the estimates made in the Annual Budgetary Law.

The fact is that these works make use of data from only one type of dataset (a federative entity). If an attempt is made to extrapolate the search for information and to carry out comparisons and analyses between various kinds of federal bodies, a number of problems will arise over the problem of the heterogeneity of the data integration. For example, education is a domain of the “function” value, an item that classifies the area covered by the public sector in which an expense is incurred. Imagine the following situation: an experienced policymaker wants to make a comparison of how much is spent on education in two cities in different States. In Brazil, before this amount can be calculated, it is necessary to know the total expenditure for education incurred by the Federal, State and Municipal governments for each city and this requires having access to five different datasets. Figure 1 illustrates this situation:

![Figure 1. Heterogeneity data on localization](image)

This heterogeneity configures the first problem: the data are not found in the same database system or directory of data archives. Another problems such as a different lexis in the header data or in the value domain datum can be observed by, for example, comparing the datasets regarding the municipality of São Paulo, the government of the State of São Paulo and the direct expenditure of the Federal Government. In the “Function" field, this has the nomenclature “Ds_Function" in the municipality of São Paulo, “FUNCTION NAME” in the government of the State of São Paulo and “ds_function” in the federal government.

As has been seen, work on the use of open government data is being carried out but so far, no work has been found that supplies the data in a clear and comprehensive way that can support and assist investigations into public information within a single information system for researchers of public policies or even the ordinary citizen. this study seeks to explore the matter and fill this gap by setting up a unique data repository on the basis of data originating from different data sources, nomenclatures, data domains and schemes in various government transparency portals of budgetary execution in Brazil.

---

1 http://dados.gov.br/dados-abertos/
2 http://www.opengovdata.org/
3 http://cuidando.org.br/
**Objectives**

The overall objective of this study is to make data from transparency portals available in a single place and in a standardized way through an information system of the Data Warehouse kind. This standardization seeks to enable the automated consumption of data to take place with less difficulty on the part of the developers of software, as well as to allow analytical queries to be made into the data of the public budget in Brazil.

**Specific aim**

The specific aim is to create an integrated data scheme where the dimensions will intersect in the same way at all the bases and thus enable a database to be formed which makes it easier to obtain information about how the budgetary execution of data can be carried out in Brazil with regard to income and expenditure.

As proof of concept the scope of this study is restricted to the data of the Federal Government, State Government and municipalities of the State of Sao Paulo, including the capital – Sao Paulo.

**Related Works**

Several countries around the world are opening up their data as a means of giving rise to applications that can assist citizens, researchers (into science and policymaking), and public managers, among others. Attention is drawn here to four studies on which the principles of this work were initially based: two in the international and two in the national sphere.

In the international domain, the Midas project (Sala et al, 2010) created an architecture for centralizing the budgetary data originating from several transparency portals. The project is based on information data structured for members of the United States Congress, electoral districts, federal agencies, government suppliers and the expenditure incurred by each. It can take the form of a federal contract and funds that are earmarked for expenditure on particular schemes or which are aimed at specific tax exemptions.

Other government works envisage the integration of data in the sphere of the Federal Government (Santos, 2011) – SIC (Federal Government Cost Information System), although this is only concerned with government management (Santos, 2011). Following this same line of reasoning, (Brito, 2014) puts forward integrated information systems as a better way of showing the data sources of heterogeneous systems such as “My National Congress” which integrates open data sources with regard to federal deputies and senators.

**Methodology**

The defined objective was achieved by initially finding the data sources which make the datasets available for each federal body within the scope of the study. Table 1 shows the location of data sources on the internet used for each federal body for income and expenditure:

<table>
<thead>
<tr>
<th>Federal Body</th>
<th>Location of datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Government</td>
<td>Transparency Portal of the Comptroller-General of the Union of the Federal Government 4</td>
</tr>
<tr>
<td>Sao Paulo State’s Government</td>
<td>Portal of the Treasury Dept 5</td>
</tr>
</tbody>
</table>

[^5]: [http://www.fazenda.sp.gov.br/download/default.shtm](http://www.fazenda.sp.gov.br/download/default.shtm)
This study was carried out in seven stages:

- Definition of a conceptual scheme for data on income and expenditure;
- Definition of the logical schemes of federal bodies and an integrated model;
- Setting up a process for retrieving data in the ETL system (Extract, Transform and Load);
- Setting up a process for transforming data from the ETL system;
- Setting up a loading process from data in the ETL system;
- Configuration of the OLAP tool for the visualization and analysis of the data;
- Validation of the prototype from the Data Warehouse System of budgetary execution.

The first stage was carried out to contextualize how budgets are classified in Brazil. The basis for undertaking this was the Public Sector Budgeting Technical Reference Manual of 2014. Figures 2 and 3 show the conceptual scheme for income and expenditure defined in the federal sphere by the MTO:

![Conceptual scheme: Expenditure and Revenue](image)

On the basis of the conceptual scheme outlined above, the second stage in the development was to identify the intersection of the fields in the databases made available for each federal body, as well as the fields that were missing and extra fields (which are not defined in the MTO, although they are available in the datasets). Table 2 indicates a color legend to the rules used in defining the logical schema of each federal body (Figure 4); and an integrated scheme formed by the intersection of the common fields of all federal bodies to be analyzed together (Figure 5).

Table 1. Location of the federal bodies datasets on the internet

<table>
<thead>
<tr>
<th>Sao Paulo State’s Municipalities</th>
<th>Citizen’s Portal of the Audit Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sao Paulo Municipality</td>
<td>Portal of the Secretariat for Planning, Budgeting and Management</td>
</tr>
</tbody>
</table>

6 http://www.portaldocidadao.tce.sp.gov.br/downloads-e-api
7 http://sempla.prefeitura.sp.gov.br/orcamento/execucao.html
<table>
<thead>
<tr>
<th>color</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>fields found in the intersection</td>
</tr>
<tr>
<td>Red</td>
<td>fields that are missing</td>
</tr>
<tr>
<td>Yellow</td>
<td>extra fields</td>
</tr>
</tbody>
</table>

Table 2. Color legend to the rules used in defining the logical schema

Figure 3. Logical Scheme example: Federal Government expenditure

Figure 4. Integrated scheme for expenditure
Once the logical schemes had been defined, the ETL process was begun. Figure 7 defines how these stages were interlinked in an integrated system:

![Architecture of the integrated system](image)

**Figure 5. Architecture of the integrated system**

The third stage began with the first phase of the ETL process: retrieval. As there is no standardization of the file path or its nomenclature and format to make available these data in the portals, a retrieval process was created for each federal body. Table 3 indicates the original format of the data in each transparency portal:

<table>
<thead>
<tr>
<th>Federal Body</th>
<th>format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Government</td>
<td>Zip of CSV file – URL Download</td>
</tr>
<tr>
<td>Sao Paulo State’s Government</td>
<td>Zip of CSV file – URL Download</td>
</tr>
<tr>
<td>Sao Paulo State’s Municipalities</td>
<td>Zip of CSV file – URL Download; XML and JSON – API</td>
</tr>
<tr>
<td>Sao Paulo Municipality</td>
<td>XLS and ODS – URL Download</td>
</tr>
</tbody>
</table>

**Table 3. Original format of the data in each transparency portal**

After the stage of the retrieval of the datasets, these underwent a transformation’s process before being loaded in the system (the fourth development stage). Figure 6 defines the flow of this process:

![Transformation’s flow process](image)

**Figure 6. Transformation’s flow process**

After the transformation of the data, the task of loading the data into the system was carried out (fifth stage of development). The loading process was undertaken in two stages: the first involved feeding the data marts of revenue and expenditure of each federal body; after this stage, the task was carried out of loading the integrated scheme through the already defined areas in the data marts.

The sixth stage of development involved the configuration of an OLAP tool for the visualization and exporting of the data. The tool chosen was the Saiku plugin which operates in a Pentaho BI Server suite, from the Pentaho community®. This tool will be used in the next stage of the methodology.
The seventh stage of development entailed validating the data in the system. The validation involved determining if the sum of the figures that the IDEO system recorded corresponded with the amount recorded by the inspecting bodies. In this stage of the validation, the data were validated by figures found in the portal of the National Treasury (for the data marts referring to data from the federal government and the State of São Paulo); and from the audit office of São Paulo (for the municipalities of São Paulo). Two sets of data were not validated.

The second part consisted of a validation of the data reports through queries made by a specialist in the area of public data. Tables 4, 5 and 6 show the results of the data consultation query in two different spheres and the results of the results of the integrated query system (Table 7). A serious difficulty was encountered in this example because of a lack of data. The data used for the municipality of São Paulo only made available the classification of the “function” from 2013 onwards.

<table>
<thead>
<tr>
<th>SP Government</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds_function</td>
<td>2010</td>
</tr>
</tbody>
</table>

Table 4. Expenses on education – Expenditure of the Government of the State of São Paulo

<table>
<thead>
<tr>
<th>SP Municipalities</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds_function</td>
<td>2010</td>
</tr>
</tbody>
</table>

Table 5. Expenses on education – Expenditure of the Municipalities of the State of São Paulo

<table>
<thead>
<tr>
<th>SP Municipality</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds_function</td>
<td>2013</td>
</tr>
<tr>
<td>Education</td>
<td>R$7.727.352.890,68</td>
</tr>
</tbody>
</table>

Table 6. Expenses on education – Expenditure of the Municipality of São Paulo

<table>
<thead>
<tr>
<th>State of SP + Municipalities</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds_function</td>
<td>2010</td>
</tr>
<tr>
<td>Education</td>
<td>R$33.867.335.834,70</td>
</tr>
</tbody>
</table>

Table 7. Expenses on education – Government of the State of SP and Municipalities

Conclusion

The aim of this study has been to show how an integrated data repository of the public budget can be formed that originates from different government sources. This takes into account the fact that Brazil is a country that is made up of a federal power, 26 States and 5,500 municipalities. As a proof of concept, data related to income and expenditure over the last 5 years. Were collected from the federal government, one
State government and 645 municipalities. This made it possible to carry out a preliminary study of the particular budgetary concepts of Brazil, as well as the laws which require transparent government data and the basic principles of open government data. Moreover, through a definition of database schemes, this allowed an important study to be conducted of the data schemes already published in the portals. This enabled further projects to be carried out which opened up a new line of inquiry, as well as finding solutions to mapping problems in distributed heterogeneous networks and correctable failures. It was also found that despite the existence of a concept and law of transparency, there still exist gaps resulting from the non-availability of data in the portals, which makes it impossible to supply certain information and also means that there is a lack of transparency in certain situations. The final product – a loaded database – has still not been completed, although it can be confirmed that this can be attached to an OLAP system for the visualization of information and exporting of data and has already produced information in a simpler form for further analysis.

Thus, this work can be defined as a prototype for a future project which seeks to build a transparency portal which standardizes and unifies the access of the general public, inasmuch as it makes the automatic retrieval of data easier and this can give rise to computational applications in new digital information contexts.

Bibliographical References


