CandiDATA: an enhanced dataset for data analysis of elections in Brazil from 1945 to 2020

Felipe F. Vasconcelos[1], João V. S. Tavares[1], Matheus G. S. Oliveira[1]
Fabio J. Coutinho[1], João Paulo Clarindo[2]

[1] Universidade Federal de Alagoas, Brazil
{ffv,jvst,mgso,fabio}@ic.ufal.br
[2] Universidade de São Paulo, Brazil
{jpcsantos}@usp.br

Abstract. The Brazilian Superior Electoral Court (TSE) keeps data on elections that have taken place in Brazil since 1933. These data constitute an important collection serving as a reference for works in several research areas. However, this collection is not fully exploited due to some problems, such as missing and non-standard data, making analysis and integration with external databases difficult. Previous works built limited datasets and tools because of these problems as they only include data since the 1998 election, disregarding the election years from 1945 and 1996. This work discusses the steps to create CandiDATA – a standardized and enhanced dataset from TSE data, including a toolkit of webscrapping and data visualization. CandiDATA is available in open format and covers the election period between 1945 and 2020.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Scientific databases; J.1 [**Administrative Data Processing**]: Government

Keywords: data integration, data cleaning, electoral data

1. INTRODUCTION

The democratic nations in the contemporary world are organized into different models and structures that reflect the characteristics of each society. Assessing the degree of maturity of a country's democracy is a complex process, as it needs to consider different aspects. *The Economist* magazine set up an index to examine the state of democracy in 167 countries based on five categories: electoral process and pluralism; government functioning; political participation; political culture, and civil liberties.

Brazil occupies the 49th position in this ranking with an overall score of 6.92. However, when the categories are separately checked the scores can be quite different, as when looking at the item "electoral process and pluralism" which received a score of 9.58, appearing as the second best score among all the countries evaluated. [Economist 2021]

Indeed, the Brazilian electoral process maintained by the Electoral Court has the recognition of international organizations, and is well known for the celerity in the counting of results, since the elections have been held entirely with electronic ballot boxes since the year 2000. Furthermore, in the last elections, more than half of the voters were identified by biometrics, an important factor to avoid possible fraudulent actions, such as voting on behalf of third parties who are already deceased. Therefore, it appears that the computerization process of Brazilian elections has evolved consistently since 1985, when the digital electoral register was implemented [Tribunal Superior Eleitoral 2016].

The computerization of the Brazilian electoral system boosts the process of generating and collecting data, such as candidate and voter profiles, counting results, accountability, electoral surveys, etc [TSE 2020]. In this context, the portal of the Superior Electoral Court (TSE) provides a compilation of raw election information including, for example, the results of elections that have taken place in Brazil since 1933. This material constitutes an important collection for statistical analysis and research by academics, journalists, sociologists, political scientists and other interested parties. However, the existence of different problems seriously hinders the manipulation and analysis of the data, such as: the absence of standards, the incompleteness of relevant data, inconsistencies in the documentation, the use of different formats to represent identical data, the use of little-known standards and formats, and the fragmentation of information into numerous files. These problems add more complexity to the data manipulation, making the analysis difficult, especially for users who are not computer scientists or have little knowledge of programming.

This paper extends CandiDATA [Vasconcelos et al. 2021], a dataset built to analyze data from Brazilian elections from 1945 to 2020 generated from data obtained from the TSE portal, which were subjected to standardization, cleaning and transformation. CandiDATA was enriched through data inference to fill in missing information regarding years prior to 1996. In this way, null, malformatted and inconsistent fields were removed while new fields were added, facilitating integration and assisting the decision-making process for managers, researchers and other stakeholders. CandiDATA is freely available in open format and can be read by analysis tools and programming languages for use in different areas and branches of society. This paper has the following proposed extensions to CandiDATA: The increase of existing information about candidates, covering data about their assets and their electoral accountability; the implementation of a semi-automatic tool for loading data into the dataset and the development of a data visualization tool.

This document is organized as follows: Section 2 discusses works related to CandiDATA; Section 3 describes the original data and reports problems found in the TSE database; Section 4 presents an overview of dataset building process; Section 5 discusses CandiDATA applicability and presents usage examples; and, finally, Section 6 concludes the paper with additional comments and future work.

2. RELATED WORK

Many works have used the *Electoral Data Repository* available on the TSE portal as a research source over the last decades. Some examples that can be mentioned are: the analytical study of the limiting factors for the implementation of the quota policy for the selection of candidates to political office in Brazil [Araújo 2010]; the analysis of the association between political finance and electoral performance in the elections for state and federal deputies in Brazil [Speck and Mancuso 2014] and a municipal level analysis of spatial and temporal patterns in the presidential election results from 1994 to 2018 [Jacintho et al. 2020].

Camargo et al. (2016) also uses the TSE data, the authors aim to find patterns in the profiles of candidates for city councilor in municipalities of Rio Grande do Sul state based on data referring to the 2012 election. Initially, the authors point out that the TSE data needed to be converted to Comma-Separated Values (CSV) files, followed by changes for attribute standardization, these are problems that CandiDATA has also adressed. Later, they applied data mining techniques to identify patterns using the J48 algorithm. The results pointed out the following factors as relevant for election to the position of councilman: the political career, age, level of education, and gender of the candidate.

Filho and Pappa (2015) developed a tool for the demographic characterization of Twitter users. From the analysis of posted messages and profile information, the user's gender, age and social class are inferred. To demonstrate the use of the tool, the authors used voter data obtained from the TSE repository in order to build a real demographic distribution to compare it with the demographic distribution built from Twitter user data.

These works serve as references of the TSE dataset usage by the scientific community. Thus, the creation of CandiDATA will help the scientific community by bringing an improved version of the original TSE dataset. Some works relate more to the implementation of CandiDATA, such as the followings:

CEPESP (2020) developed the CepespData platform in order to facilitate integrated access to databases made available by TSE in its repository. The data maintained in the platform can be obtained directly from the CepespData[1] website or through Application Programming Interfaces (APIs) that can be accessed through programming languages such as R and Python. The work overcomes some problems found in the TSE repository, however, the data are restricted to the period from 1998 to 2018. In contrast, CandiDATA gathers data from 1945 to 2020, allowing for more comprehensive analyses with a historical perspective. It is worth noting that the older data have a greater number of problems such as inappropriate formats, lack of standards and relevant information. Considering the data visualization, there are some interesting initiatives, TSE has a dashboard called Electoral Statistics[2], providing information about candidates, election results and data crossing. *Poder360* newspaper developed a website[3] which gathers information about the candidates since 1998 elections such as the number of votes, declaration of assets, political party, etc. However, both platforms have limitations in their representations. TSE site brings a business intelligence approach, but is limited in the amount of data, ranging only from 2014 to 2020 and not showing important information such as the amount of votes of each candidate. Poder360 website despite a greater temporal scope of information, from 1998 to 2018, it does not use statistical tools to analyze of this data, being just a research source. CandiDATA dashboard includes electoral data from 1945 to 2020 and provides views, statistical analysis and data crossing. Table I highlights the difference between CandiDATA and other works based on three parameters: Time range; Data visualization and Data categories.

Table I: Differences between the related works

WORK	TIME RANGE	DATA VISUALIZATION	DATA CATEGORIES
CANDIDATA	1945-2020	Dashboard with Graphics	Candidates; Accountability; Assets
CEPESPDATA	1998-2018	SQL Queries	Candidates; Accountability; Assets
ELECTORAL STATISTICS	2014-2020	Business Intelligence Approach	Candidates; Accountability
PODER360	1998-2018	Candidates Personal Page	Candidates; Assets

3. TSE PORTAL DATA

CandiDATA was built from data obtained from the TSE portal[4] which are organized into different categories, such as Candidatos, Resultados, among others. These categories are divided by year, with each year subdivided by federal unit. The data available cover the electoral periods from 1933 to 2020. This fragmented organization of the data demands more effort in exploring the content. For example, it is necessary to access several parts of the site to gather information about a particular candidate.

TSE portal provides data in ZIP-formatted compressed files, which contain text files without headers using semicolon delimitation and `Latin1` encoding. However, data from elections starting in 2012 are stored in CSV files, with the same delimiter as previous years, but including headers. Considering the data about the number of votes and declaration of assets of the candidates, PDF files entitled "leia-me" are provided, which describe the data dictionary. Regarding the accountability of candidates, there is a textual file called "leioute" that describes the data dictionary.

[1] `https://cepespdata.io/`
[2] `https://www.tse.jus.br/eleicoes/estatisticas/estatisticas-eleitorais`
[3] `https://eleicoes.poder360.com.br/`
[4] `https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1`

3.1 Problems found in TSE data

As mentioned earlier, the analysis of the data published on the TSE portal requires considerable effort due to the presence of several problems such as: inconsistency in data and documentation, absence of known standards and formats.

The main problems found and solved in the TSE data during the construction of CadiDATA are described in detail below.

—The **format of the data files** is irregular, since until the 2010 election the files are in TXT format without header, while from 2012 the data files are in CSV format with header.

—**Inconsistency in the description of fields in the dictionaries**, for example, there are different names for the same attribute, in addition, there is a lack of standardization of the dictionaries of the same category, the most critical case being Accountability, where there is a different dictionary for each election year, and there are also years without data dictionary such as 2014.

—Representation of the contents in `latin1`, which can cause errors when loading these data because an international standard is not considered.

—**Date type fields have standardization failures**, including distinct formats such as `dd/MM/yyyy`, `ddMMyy` ou `dd/MM/yy`.

—Attributes related to **municipal codes** do not use the city identification standard of the Brazilian Institute of Geography and Statistics (IBGE).

—**Lack of standardization in the voter ID number**, since the current standard is 12 digits and there are data represented by old standards with less than 12 digits.

—**Absence of the gender field** for candidates in the years 1945 to 1994.

—The **names of the attributes are not standardized**; each election year presents a different format for the attributes.

—**There are different representations of null fields**: `#NULO#` and `#NE#` to *string*; and `-1` e `-3` to integer.

The problems mentioned above were addressed in the three CandiDATA modules, which will be presented in Section 4.

4. BUILDING THE DATASET

CandiDATA dataset comprises three parts related to elections: *Voting*, which contains data about the candidates with their respective number of votes in the election race; *Candidate Assets*, which maintains information about the candidate's list of assets and *Accountability*, which contains data about the accountability statements, including revenues and expenses of each candidate's electoral campaign. Figure 1 illustrates the stages of CandiDATA developing, organized into three modules: (i) standardization – responsible for standardizing the data regarding null fields, dates, voter ID number, etc; (ii) transformation – responsible for inserting new information, such as converting city codes to the official IBGE standard, adding the Brazilian Code of Occupations (CBO) and the gender of candidates from its inference, as well as information obtained from data crossing; and (iii) consolidation – step in which the data is arranged in the way described in the dictionary, both in CSV format and JSON format.

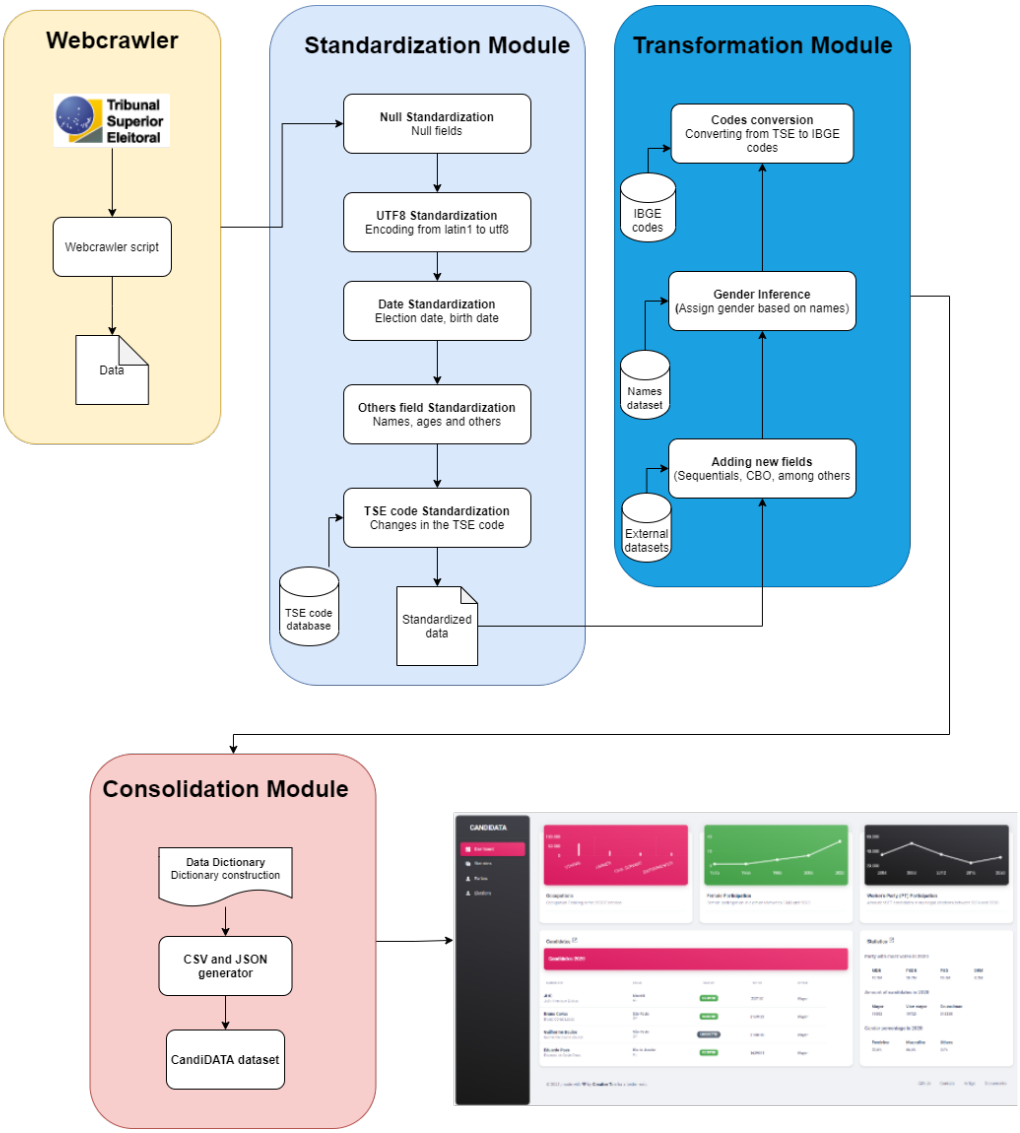

Fig. 1: CandiDATA dataset building execution flow

These modules were developed in the Python language due to the ease of manipulating data and CSV files and the existence of a large collection of consolidated libraries in this area, such as the NumPy[5] library and the Pandas[6] library for data analysis. In the next sections, the procedures performed in each module are discussed.

4.1 TSE Data Collection - Webcrawler

In order to collect data from TSE website, a webcrawler was developed using Python language with the Selenium library[7]. This process works in a semi-automated way, through the initialization of a script that aims to standardize and speed up the task of obtaining the files available on the TSE website by

[5] `https://numpy.org/`
[6] `https://pandas.pydata.org/`
[7] `https://www.selenium.dev/`

having as the input the electoral year and as the output the data from that electoral year. Thus, the webcrawler is responsible for going through each section of the TSE portal, verifying the existence of new files and then downloading all the files needed to update the CandiDATA dataset. The main functionalities of the data collection include detecting the existence of new TSE data, monitoring the download of files of interest to CandiDATA, and logging the update performed in order to notify the dataset users.

CandiDATA data files are organized into three folders that represent the dataset groups: voting, accountability, and candidate assets. Each folder has a hierarchy of subdirectories divided into election years. In more recent electoral years, information is organized in 27 files, representing each federative entity, while older elections have only 8 data files. This difference in the number of files is due to the availability of data on the TSE website. The TSE does not give explanations about the non-existence of data from the other 19 states in Brazil.

Considering the raw data from the TSE, there was a redundancy of data referring to the voting of candidates, given that the result is divided by electoral zones, where the same candidate has its data repeated several times, according to the electoral zone. To reduce this redundancy, CandiDATA provides a view that contains the cumulative result of the number of votes received by candidates. Thus, considering the 2020 elections, the number of lines was reduced from around 2.5 million to an average of 15 thousand lines. Therefore, the CandiDATA dataset occupies about 40GB.

Another aspect of dataset quality is the high occurrence of null values for some fields. This fact can be explained by the inclusion of new data over time, during the evolution of the electoral process, and the integrated view of the CandiDATA dataset, which maintains a single data dictionary. For example, in elections prior to 1994, there was no data referring to the field SQ_CANDIDATO, which represents a sequence generated for each candidate in the electoral system. The same occurs for other fields such as SQ_COLIGACAO and NM_COLIGACAO in election years in which there were no party coalitions.

4.2 Standardization Module

As described in Section 3.1, problems related to the lack of standardization were found in the TSE data. Therefore, the initial stage of building CandiDATA consists of developing tasks for standardizing relevant data, such as:

—**Voter ID Number**: All voter ID numbers have been standardized to the current twelve-digit format by adding zeros to the left of the number.

—**Null fields**: Null fields have been standardized by assigning `NULL` values for *strings* and `-1` for *integers*.

—**Date fields**: Dates such as birth and election date have been standardized in the format `yyyy-MM-dd`.

—**Text fields**: Names and other string-type fields have been encoded to the UTF-8 specification using the Unidecode[8] library.

4.3 Transformation Module

In the transformation module, tasks are performed to correct inconsistent information and add missing data to the TSE database. Considering the voting data, we highlight the inclusion of the IBGE municipal identification code and the code referring to the Brazilian Classification of Occupations

[8] `https://pypi.org/project/Unidecode/`

(CBO). In addition, another relevant action was the inference of the gender of the candidates from their names.

The conversion process of the municipal codes was done with the help of a TSE/IBGE code conversion dataset[9]. This conversion aimed a better standardization of CandiDATA, since the TSE database uses its own code to identify each Brazilian municipality. The use of the IBGE city code standard facilitates a possible integration of CandiDATA with external sources, such as the social and economic databases made available by IBGE.

Considering the professions of the candidates, a *dataset*[10] was used, which contains data relating to the Brazilian Classification of Occupation (CBO). These codes, standardized by the Ministry of Labor [MTE 2020], are used in different contexts, for example, in databases related to public health [Clarindo et al. 2019], where CBO codes are used to identify the occupation of patients. The conversion to CBO can generate ambiguity for some cases, where these ambiguities are generated due to the differences between the professions dataset used by the TSE and the dataset used by CBO, which result in not always existing a direct transformation. As an example, the profession classified as "State public servant" in the TSE, does not have a CBO code that directly correlates to it, so in this and other similar cases a null value is assigned.

An important issue verified in the TSE data was the lack of gender of the candidates in electoral years prior to 1994. Thus, the CandiDATA construction performed the gender inference from the use of a public name base[11], comparing the first name with the gender listed in the name dataset. The inference of genders is based on the probability of a name belonging to a gender, given that the dataset contains information about the gender frequency of each name and its variations. In cases of names that can be used by both genders, the classification will be done according to the probability. In cases where there was no equivalence of names, about 0.20%, a null value was assigned for the gender of the candidate in question. For candidates of foreign origin, there may be error in inferring gender from the Brazilian base of names, due to the variation that occurs from one country to another. It was not possible to find a generalist solution to the problem, since the field describing the nationality of candidates does not exist in the election years between 1945 and 1990.

In the part related to the candidates' assets, the only information about the candidate, available in the TSE data, was his sequential electoral code. In CandiDATA, the NM_CANDIDATO field was added, which contains the candidate's name, allowing another form of identification and facilitating dataset manipulation. This information was obtained by cross-referencing data from the Voting segment with the Candidate Assets segment.

4.4 Consolidation Module

Finally, the consolidation module has as its main function the transformation of the final file into CSV format, which is an open format, regulated by the *Internet Engineering Task Force* (IEFT), consisting of a text file with tabular format, that uses commas as delimiters [Shafranovich 2005] and in JSON (*JavaScript Object Notation*) format, an open format that is easy to read [Bray 2017]. The JSON and CSV formats complement each other: while CSV is usable by many DBMSs, JSON is used in APIs and NoSQL DBMSs [Drapeau 2018].

In TSE data, a considerable part of CSV files do not have headers, making them difficult to manipulate. For each election year, the construction of the CandiDATA CSV files has inserted headers containing the standardized fields in a single format.

In relation to data dictionaries, from the data dictionaries provided by TSE, new standardized

[9] https://github.com/betafcc/Municipios-Brasileiros-TSE
[10] https://github.com/datasets-br/cbo
[11] https://brasil.io/dataset/genero-nomes/nomes/

dictionaries were built for CandiDATA, including the new fields added to ensure compatibility between all the electoral years available in the dataset. The most recent TSE version was used as a reference, in order to facilitate updating the data in future elections.

The data dictionary of the Voting segment is presented in Table II,which includes new attributes that did not exist in the original dictionary from TSE, and were created by CandiDATA, such as NR_IDADE_DATA_POSSE (age of the candidate), DT_NASCIMENTO (candidate's date of birth), NR_TITULO_ELEITORAL_CANDIDATO (candidate's voter ID number), CD_GENERO (candidate's gender code), DS_GENERO (candidate's gender), CD_CBO (candidate's occupation code), CD_MUN_IBGE (municipality code based on IBGE), and the standardized fields described in Section 4.3.

Table II: CandiDATA data dictionary, where fields highlighted in red refer to fields that do not exist in the TSE original databases and were added by CandiDATA.

VARIABLE	DESCRIPTION
DT_GERACAO	Data extraction date for the files by the Superior Electoral Cort.
HH_GERACAO	Data extraction hour for the files by the Superior Electoral Cort.
ANO_ELEICAO	Election year.
CD_TIPO_ELEICAO	Election type code.
NM_TIPO_ELEICAO	Election type name.
NR_TURNO	Elections round number.
CD_ELEICAO	Election code..
DS_ELEICAO	Election description.
DT_ELEICAO	Elections date.
TP_ABRANGENCIA	Elections coverage (Federal or municipal).
SG_UF	Acronym of the Federation Unit in which the election took place.
SG_UE	Acronym of the District Unit of the candidate.
NM_UE	Name of the District Unit of the candidate.
CD_MUNICIPIO	TSE code of the city in which the election took place.
NM_MUNICIPIO	Name of the city in which the election took place.
NR_ZONA	Electoral Unit code of the city in which the election took place.
CD_CARGO	Candidate office code.
DS_CARGO	Candidate office description.
NR_IDADE_DATA_POSSE	Candidates age.
DT_NASCIMENTO	Date of birth
SQ_CANDIDATO	Candidate sequential code generated internally by the electoral systems.
NR_CANDIDATO	Candidates number
NR_TITULO_ELEITORAL_CANDIDATO	Candidate voter id.
NM_CANDIDATO	Candidates full name.
NM_URNA_CANDIDATO	Candidates name in the eletronic voting machine.
NM_SOCIAL_CANDIDATO	Candidates social name.
CD_GENERO	Candidates gender code.
DS_GENERO	Candidates gender.
CD_SITUACAO_CANDIDATURA	Candidate registration status code.
DS_SITUACAO_CANDIDATURA	Candidate registration status description.
CD_DETALHE_SITUACAO_CAND	Candidate registration details code.
DS_DETALHE_SITUACAO_CAND	Candidate registration details description.
TP_AGREMIACAO	Coalition type.
NR_PARTIDO	Candidate's party number.
SG_PARTIDO	Candidate's party Acronym.
NM_PARTIDO	Candidate's party name.
SQ_COLIGACAO	Coalition sequential code generated internally by the electoral systems.
NM_COLIGACAO	Candidates coalition name.
DS_COMPOSICAO_COLIGACAO	Candidates coalition composition
CD_SIT_TOT_TURNO	Candidate round status code.
DS_SIT_TOT_TURNO	Candidate round status description.
QT_VOTOS_NOMINAIS	Candidate vote amount
CD_CBO	Candidates occupation code in the Brazilian Occupation Classification.
CD_MUN_IBGE	City TSE code

In the Candidates Assets segment a new field was added: NM_CANDIDATO, referring to the candidate's name, since the TSE only uses the candidate's sequential to identify him or her, which can cause more difficulties in viewing this data. In addition, DT_ULTIMA_ATUALIZACAO and HH_ULTIMA_ATUALIZACAO referring to the last update of the candidate's assets were excluded since the fields DT_GERACAO

and HH_GERACAO already inform the last update made by the TSE. The dictionary of the Candidate Assets segment is presented in Table III.

Table III: CandiDATA data dictionary referring to Candidate Assets, where the fields highlighted in red refer to fields that do not exist in the original databases

VARIABLE	DESCRIPTION
DT_GERACAO	Data extraction date for the files by the Superior Electoral Cort.
HH_GERACAO	Data extraction hour for the files by the Superior Electoral Cort.
ANO_ELEICAO	Election year.
CD_TIPO_ELEICAO	Election type code.
NM_TIPO_ELEICAO	Election type name.
CD_ELEICAO	Election code..
DS_ELEICAO	Election description.
DT_ELEICAO	Elections date.
SG_UF	Acronym of the Federation Unit in which the election took place.
SG_UE	Acronym of the District Unit of the candidate.
NM_UE	Name of the District Unit of the candidate.
NR_ORDEM_CANDIDATO	Order number of the candidates asset.
CD_TIPO_BEM_CANDIDATO	Candidate assets type code.
DS_TIPO_BEM_CANDIDATO	Candidate assets type description.
DS_BEM_CANDIDATO	Candidate asset description.
VR_BEM_CANDIDATO	Candidate asset value.
NM_CANDIDATO	Candidates name.

Considering the Accountability segment, there was a strong inconsistency in the data dictionaries made available by the TSE over the different election years. In 2002, there are only twelve fields, while in 2020 there are 57 fields. Furthermore, no data dictionary was made available in the 2014 election year. Therefore, the data dictionary built for CandiDATA has a final reduced number of fields, maintaining a simpler and more homogeneous version, which lists the candidates' income and expenses as described in Table IV.

Table IV: CandiDATA data dictionary referring to accountability of financial incomes and expenses, where the fields highlighted in red refer to fields that do not exist in the original databases.

VARIABLE	DESCRIPTION
SQ_CANDIDATO	Candidate sequential code generated internally by the electoral systems.
SG_UF	Acronym of the Federation Unit in which the election took place.
SG_PARTIDO	Candidate's party Acronym.
NM_CANDIDATO	Candidates full name.
DT_RECEITA	Financial income date.
NR_CPF_CNPJ_DOADOR	Donor document number.
SG_UF_DOADOR	Donor Federation Unit acronym.
NM_DOADOR	Donor full name.
VR_RECEITA	Financial income value.
TP_RECURSO	Type of financial resource.
SQ_RECEITA	Financial income sequential code generated internally by the electoral systems.
DS_RECEITA	Financial income description.

4.5 Dashboard

As discussed in Section 2, the data visualization options on the TSE website are limited to data up to 2014. In CandiDATA, a dashboard was created in order to present data referring to electoral years from 1945 to 2020. This tool allows for cross-referencing information and statistical analysis.

The main objective of the dashboard is to explore the capacity of the CandiDATA dataset, demonstrating how researchers and interested parties can view pre-assembled queries and learn about the capacity to generate information using the dataset. At the current stage of implementation, the data visualization tool is divided into four parts as described below.

—**Main**: the main view will encompass a bit of all parts of the dashboard, with graphs, figures and candidate displays, to serve as an overview of the dataset's application capabilities. It is represented in Figure 2.

—**Statistics**: it contains an exclusive page for statistical information, with graphs encompassing the ranking of Occupations, the most voted political parties, the parties with more candidates, among other information, presenting data about several electoral years such as 1945, 1990 and 2020.

—**Political Parties**: it presents the list of candidates who ran in that year's elections for the selected party, including information such as the candidate's full name and number, as well as the party's logo.

—**Candidates**: It displays information on candidates for the election year, such as the number of votes, status, information about their accountability, as well as general information such as statements of the largest donors, the categories with the highest expenses, among other information.

The dashboard was developed in HTML, CSS and JavaScript, using an open-source template based on Bootstrap and ChartJS, which was modified for CandiDATA's needs. Queries were made in Python and from the results JSON files were generated, which are hosted in the Firebase Realtime Database.

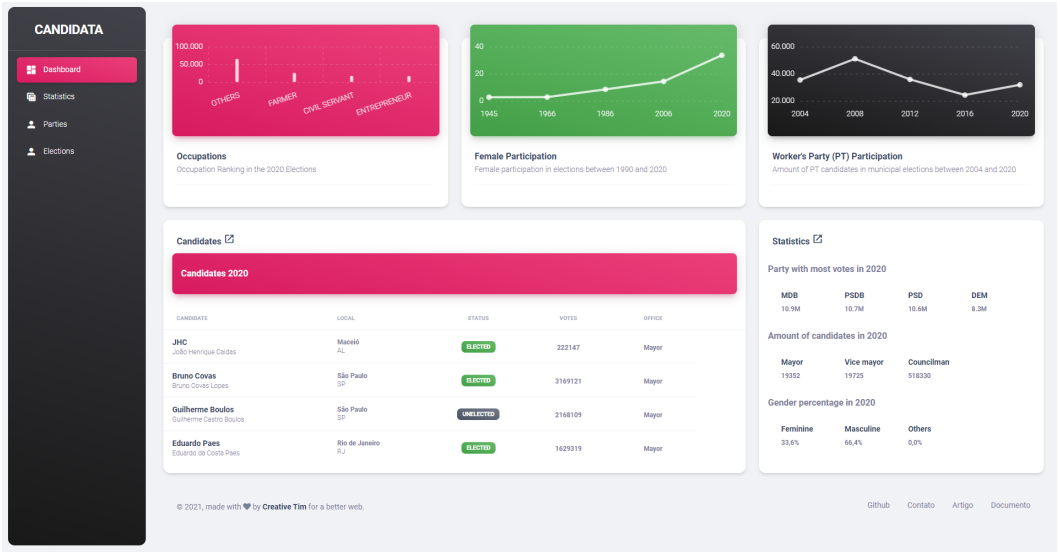

Fig. 2: Dashboard main view

5. CANDIDATA APPLICATIONS

An important differential of CandiDATA is the breadth of the period in which information is made available, the inclusion of additional fields referring to gender, occupations and electoral titles of candidates, and the standardization of data dictionaries. In addition, another relevant aspect is the feasibility of integration with external databases through the use of standardized fields. Therefore, CandiDATA can be used by managers and researchers interested in the Brazilian elections, as well as by IT professionals seeking standardized open data for different purposes.

As mentioned in Section 2, there are works that address numerous aspects about the Brazilian elections, such as female participation, limited representativeness and lack of transparency. In view of this,

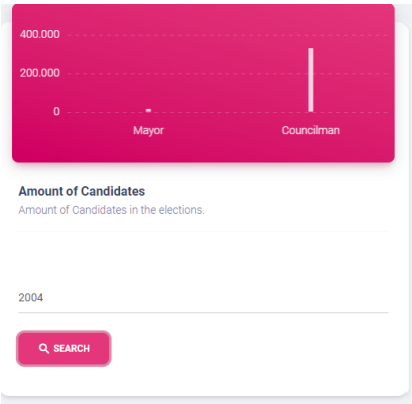

Fig. 3: Amount of candidates per position

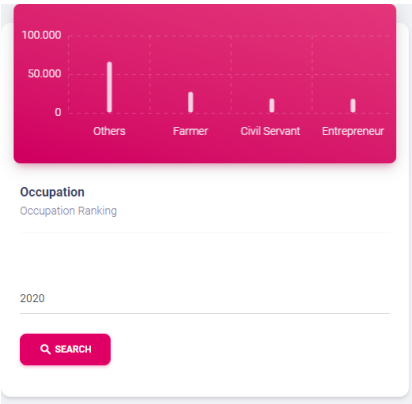

Fig. 4: Occupation Ranking

the CandiDATA dataset has the potential to help promote several analyses with historical perspective, such as, for example, the construction of graphs showing the growth of women's participation in elections over the years. In addition, the dataset opens the possibility to establish comparisons and correlations between different characteristics of the candidates such as gender, education rate, profession, age group, party affiliation, etc.

Thus, with the help of CandiDATA you can perform various queries that generate relevant information for society. As a demonstration, the sections 5.1, 5.2, 5.3 and 5.4 present examples of the use of dataset.

The dashboard presented in this work also aims to demonstrate to the users some use cases for the CandiDATA data, such as the number of candidates per position, represented in Figure 3, ranking of professions in the elections, represented in Figure 4, and others.

5.1 Women's participation in politics

Using the CandiDATA, it was possible to analyze the female participation in the electoral process over the last few years. Figure 5 demonstrates the evolution of women's participation as candidates in elections each year. There has been a significant increase in female participation from 2008 onwards.

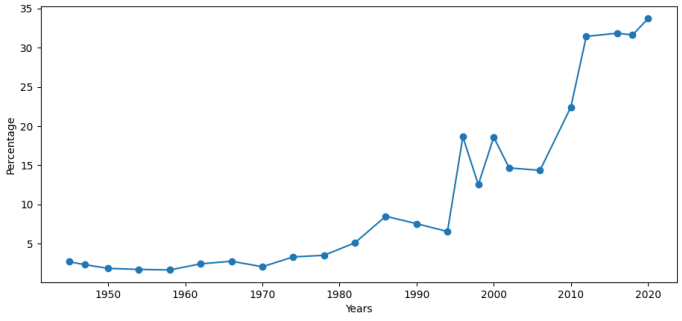

Fig. 5: Percentage of female participation in elections

5.2 Occupations with higher participation in 2020 election

In Figure 6 it is possible to verify the professions with the highest participation among the candidates of the 2020 election. The profile of farmer was the most popular among candidates in the 2020 municipal elections.

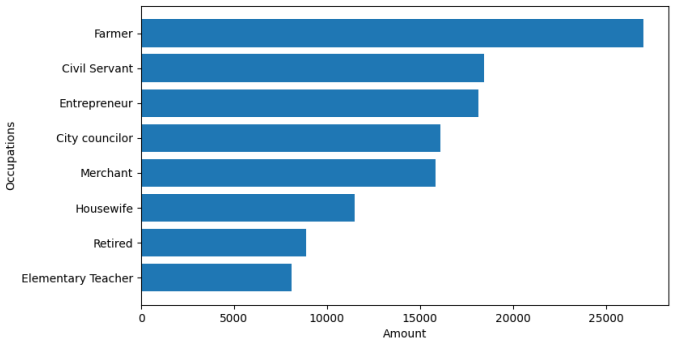

Fig. 6: Candidate occupations in the 2020 elections

5.3 Political parties with the most candidates between 2000 and 2020

Figure 7 shows the number of candidates, in the last three electoral years, from the parties of the last five presidents of Brazil, being the parties: PSL, PT, MDB, PSDB. One can see the rise of the PT in mid-2002 and 2008, possibly motivated by the two presidential elections won by the party, in 2002 and 2006. It is also possible to analyze the rise of the PSL, the party that won the presidential race in 2018, which managed to increase in ten times the number of candidates between 2000 and 2020, going from around 1100 to more than 11 thousand candidates, respectively. It appears that the MDB, despite losing popularity, as well as the other traditional parties, still maintains its leadership in number of candidates.

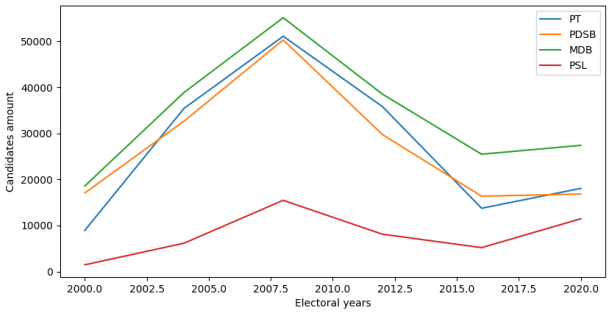

Fig. 7: Most candidates parties between 2000 and 2020

5.4 Candidate assets in 2018

In this section, we present the visualization of the wealth belonging to the candidates. The graph 8 demonstrates the five largest averages, in a million, of these assets in the 2018 presidential election, being the states: Tocantins, Mato Grosso, Acre, São Paulo, and the Federal District. While Graph 9

represents the five largest total sums of the candidates' fortunes, in trillions. One can see the great disparity between São Paulo, the richest state in the country, and the other states, with a difference of more than 3 trillion to the second place.

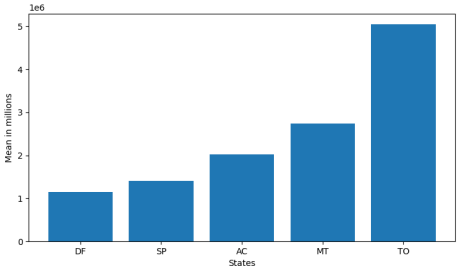

Fig. 8: Five largest equity averages in Brazil

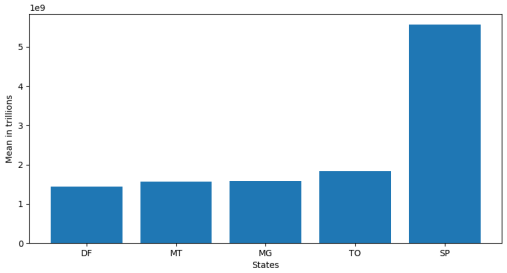

Fig. 9: Five largest sums of assets in Brazil

6. CONCLUSION

This paper presented the dataset CandiDATA, which makes available a dataset on Brazilian elections that occurred between 1945 and 2020, a webcrawler for downloading the files and a panel for visualizing the data. The data have been standardized and include additional fields that assist researchers and managers in analysis and possible integration with external bases. Some limitations are found in CandiDATA such as, for example, cases where it was not possible to infer the gender of the candidate from the base of names, therefore assigning the null value for these situations, besides the loss of some fields given the large temporal scope and the amount of dictionary changes, which led to some compatibility problems between fields. There were also cases of data typing errors by the TSE, so there is a difficulty in correcting this data, since the official source is the TSE itself.

In the future, we intended to expand the CandiDATA to include another segment, addressing information about the voters, such as education level, age groups, geopolitical information, among others. It is also intended to extend the dashboard, adding new views, as well as offering functions such as a search for an exclusive candidate.

The dataset is available for download at `https://github.com/candidata/candidata_dataset`, and is divided into 3 folders, containing each section of the dataset, which is divided into election years. In this repository are also available the data dictionary containing the description of each field. The file names are as follows: "votacao_candidato_munzona_YEAR_UF", "bem_candidato_YEAR_UF", "receitas_candidatos_YEAR_UF", "despesas_candidatos_YEAR_UF". The files are available in both CSV and JSON formats in the repository. The dashboard is available at `https://candidata.github.io`

REFERENCES

Araújo, C. The limits of women's quotas in brazil. *IDS Bulletin* 41 (5): 17–24, 2010.

Bray, T. The JavaScript Object Notation (JSON) Data Interchange Format. RFC 8259, RFC Editor, 2017.

Camargo, A., Silva, R., Amaral, E., Heinen, M., and Pereira, F. Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do Rio Grande do Sul. *Brazilian Journal of Applied Computing* 8 (1): 64–73, abr., 2016.

CEPESP, F. Cepespdata - political database. `https://empregabrasil.mte.gov.br/76/cbo/`, 2020. [Online; access in aug. 12].

Clarindo, J. P., Fontes, W., and Coutinho, F. QualiSUS: um dataset sobre dados da Saúde Pública no Brasil. In *XXXIV brazilian symposium on Databases: Dataset Showcase Workshop, SBBD 2019 Companion*. SBC, Fortaleza, CE, Brazil, October 7-10, 2019, pp. 418–428, 2019. in Portuguese.

Drapeau, M. The State of CSV and JSON. https://medium.com/@martindrapeau/the-state-of-csv-and-json-d97d1486333, 2018. [Online; access em jul. 19].

Economist, T. Global democracy has a very bad year, 2021.

Filho, R. M., Almeida, J., and Pappa, G. Pesquisa eleitoral em redes sociais: Inclusão da análise de novas dimensões. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*. SBC, Porto Alegre, RS, Brasil, pp. 164–175, 2014.

Jacintho, L. H., da Silva, T., Parmezan, A., and Batista, G. Brazilian presidential elections: Analysing voting patterns in time and space using a simple data science pipeline. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 217–224, 2020.

MTE. Brazilian classification of occupations. https://empregabrasil.mte.gov.br/76/cbo/, 2020. [Online; access in aug. 11].

Shafranovich, Y. Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180, RFC Editor. October, 2005.

Speck, B. and Mancuso, W. A study on the impact of campaign finance, political capital and gender on electoral performance. *Brazilian Political Science Review (Online)* vol. Vol. 18, pp. P. 34–58, 04, 2014.

Tribunal Superior Eleitoral. Brazilian electronic voting machine : 20 years in favor of democracy. *Electoral Superior Court*, 2016.

TSE. Repositório de dados eleitorais. https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1, 2020. [Online; access in aug. 11].

Vasconcelos, F., Tavares, J., Ribeiro, M., Coutinho, F. J., and Clarindo, J. P. Candidata: um dataset para análise das eleições no brasil. In *Anais do III Dataset Showcase Workshop*. SBC, Porto Alegre, RS, Brasil, pp. 160–168, 2021.