Assessing the capability of open government data for process mining: a study on a Brazilian open data portal

*Gyslla*Vasconcelos[a*], *Jose*Viterbo[a], *Flavia*Bernardini[a]

[a]Institute of Computing, Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brazil, gysllav@id.uff.br, {viterbo, fcbernardini}@ic.uff.br, 0000-0001-7886-5109, 0000-0002-0339-6624, 0000-0001-8801-827X.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 26 May 2025

Abstract. The growth in Open Government Data (OGD) availability over the last decade offers significant potential for promoting transparency and improving operational efficiency in government. OGD can support the identification of inefficiencies and the extraction of process models through Process Mining (PM) applied to public service tasks. Various OGD portals, maintained by federal and subnational governments, provide access to diverse public datasets that can be explored for these purposes. Applying PM to OGD, however, requires addressing key aspects such as the tabular nature of data, variability in quality and standardization, and the frequent lack of process context. This study proposes a method to classify and evaluate OGD datasets based on their relevance and potential for process discovery. The method was applied to data from the Brazilian Federal OGD portal (dados.gov.br) to investigate whether OGD can be effectively used in PM tasks to identify processes and bottlenecks. The research involved steps of data collection, selection, and evaluation. Datasets were classified based on their potential utility and suitability for transformation into event logs. The results showed that 24% of the sampled datasets were considered relevant for PM, with 23% in dated tabular format and 1% already in event log structure. Moreover, 52% of the datasets addressed public policies, suggesting the potential of PM to reveal inefficiencies in processes that affect citizens. These findings demonstrate that, with structured evaluation criteria, PM can be effectively applied to public data. This research contributes by presenting an empirical and replicable method to support the evaluation and preparation of OGD for PM applications in digital government. The results reinforce the importance of adopting clear assessment protocols to broaden the use of OGD in understanding and improving public processes.

Keywords. Digital Government, Process Mining, Open Government Data, Public Service Optimization.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.1041

1. Introduction

In the context of digital government, Open Government Data (OGD) is a transparency policy that enables broad dissemination of information on public services such as hiring, document issuance, and certifications (Jetzek et al., 2019). This dissemination seeks to enhance transparency and support external analyses that contribute to optimizing these services. Despite its potential, studies applying Process Mining (PM) to OGD remain limited and are considered rare (Rawiro et al., 2023), revealing a gap in the literature and the need for methodologies that leverage OGD to discover and improve public service processes. PM plays a key role in public administration by enabling process discovery and operational analysis based on real activity data. Its openness and adaptability present valuable opportunities to expand its application in the public sector.

Another challenge involves citizen engagement in data analysis (Barcellos et al., 2024). Improved methods are essential to facilitate access and interpretation of OGD, enabling more meaningful participation and con-

tributing to better public service processes (Lanoue, 2020). Accessible metadata can democratize analysis and promote broader involvement, helping to harness OGD's potential to identify and address service bottlenecks. In this context, bottlenecks refer to delays, rework, or inefficiencies in public service operations revealed through the process execution flow.

Recent research has explored the application of PM in various public contexts, such as improving processing times in birth registration in Indonesia (Febrianti, 2024), and in Brazil, analyzing judicial data to enhance the efficiency of the justice system (Unger et al., 2021). It has also been used to identify gaps, bottlenecks, and redundancies in regulatory processes proposed by the Federal Executive Branch (da Costa et al., 2020). However, the application of PM on OGD faces significant challenges due to availability, data quality and suitability for such analyses, like the lack of specific context that accompanies the data.

This study aims to assess these challenges by proposing a method for evaluating OGD datasets, available on OGD portals, in order to identify which datasets can be effectively used in PM tools, particularly to discover processes and address bottlenecks in public service operations. Our main Research Question (RQ) in this work is to verify whether OGD can be used to identify processes and their bottlenecks, which we divided into two Research Questions (RQs): (RQ1) "How to evaluate the application of OGD in PM tools to discover processes and bottlenecks in the public service?", and (RQ2) "What criteria are needed to select and prepare OGD for process mining?".

Our method was applied on the Brazilian Federal Government's open data platform. We observed a significant and varied data source indicating that our method offers an approach tailored to the complexities of government data (da Costa et al., 2020; Macedo & da Silva Lemos, 2024). In this context, the research was structured into three main phases: data collection, selection, and evaluation. During the collection phase, the open data portal to be used was mapped and defined, followed by the determination of the sample, collecting datasets, organization, and classification based on relevance. In the selection phase, a preliminary selection of datasets was carried out, where ranking criteria were established, calculated, and normalized for standardized comparison, leading to the selection of the highest-ranked datasets for testing in PM tools. In the final phase, the tests were executed, and the final evaluation was conducted.

Our work contributes by proposing a categorization for the data available on Brazilian government open data portals and to the practices of PM on OGD datasets, providing steps for the evaluation and preparation of datasets that facilitate the identification of processes and operational bottlenecks. From this work, researchers and professionals can evaluate and prepare OGD for PM analyses, to identify the inherent processes within the data and detect bottlenecks, thereby enhancing the understanding of the data and the efficiency of governmental processes. This paper has been structured as follows: a background of PM to process discovery and bottlenecks identification has been presented in Section 2. Literature review in Section 3. Our Method in Section 4. In Section 5 the analysis of dados.gov.br. Finally, the conclusions and future work in Section 6. Supplementary sections such as acknowledgments and appendix follow the main content.

2. Process mining to process discovery and bottlenecks identification on OGD

PM is considered the missing link between process science and data science, bringing process perspective and behavioral insights to machine learning and data mining (der Aalst, 2016). It can be applied across domains to identify processes and bottlenecks, supporting improvement efforts. By 2025, 80% of companies are expected to use PM in at least 10% of their operations (Kerremans et al., 2023), and by 2026, 25% of global companies should adopt PM tools (Kerremans et al., 2024). Google Trends data shows increased interest in "process mining" since 2019, especially in Brazil (Google LLC, 2024).

Event data, which provide details on when, where, and under what context specific activities occur, are crucial for PM. Despite the varied sources and forms of these data—from event logs to tabular datasets—aligning them with a process perspective remains a challenge (de Murillas et al., 2019; der Aalst, 2016). This is especially true for OGD, which often lacks detailed process context. The term "data without context" refers to the absence of crucial information about data origin, collection methods, and purpose, making analysis more complex. This is a specific challenge in PM on OGD, as there is usually no access to data originators for clarifications that could enrich interpretation.

From a digital government perspective, as PM continues to advance, its application in the public sector be-

comes increasingly strategic. Since the 1950s, digital government efforts have aimed to enhance transparency, efficiency, and citizen engagement. Parks (1957) emphasized that open government and access to information should be the norm in public administration worldwide. These initiatives sought greater accountability, driven by the goal of democratizing knowledge and increasing transparency (Barcellos et al., 2022; Corrêa et al., 2017). Supported by Freedom of Information (FOI) laws, OGD portals serve as key platforms to promote transparency, improve service quality, and enable citizen participation in democratic processes (de Oliveira et al., 2024).

To enable citizen participation through OGD, data must be accessible and interpretable (Barcellos et al., 2022). Challenges such as complexity, lack of standardization, availability, and format (Rajabi et al., 2023; Vasconcelos et al., 2020) hinder the effective use of OGD. Rajabi et al. (2023) note that most OGD portals do not adhere to Linked Data standards (Berners-Lee, 2006), while Vasconcelos et al. (2020) highlight inconsistencies that make preprocessing essential. Furthermore, public data is often statistical, such as census or demographic tables (Rajabi et al., 2023). Although standardized data improves understanding and decision-making (Heumann et al., 2024), PM requires more specific characteristics to be feasible: sequential event data with case identifiers, timestamps, and activity labels (Burke et al., 2024; der Aalst, 2016; Elkoumy et al., 2023).

These structural elements (caseid, timestamp, and activity) are mandatory for datasets to be used in PM. In this work, to test the datasets identified through our method, we selected ProM, Disco, and Celonis, three widely used tools in academia and industry (de Vasconcelos et al., 2024a). ProM, an open-source tool, is known for its flexibility and rich plugin ecosystem (der Aalst, 2016); Disco, developed by Fluxicon, stands out for its intuitive interface; Celonis leads the market with support for automated data entry and complex analytics (Kerremans et al., 2024). Despite the opportunities OGD offers for process discovery, applying PM techniques remains challenging, particularly due to data inconsistency and absence of context. The next section explores how these limitations are addressed in the literature.

3. Literature Review

Recent publications indicate a growing interest in digital government topics, including OGD. This trend was verified by a search in the Scopus database using queries related to broader themes connected to this study: digital government or open data (egov), digital government in computing (egov-comp), and process mining (pm)[1]. According to the retrieved data, publications in the egov category reached 5,851 in 2024, an increase of 7% compared to 2023, reinforcing the relevance of initiatives exploring the reuse of public data to improve government processes. Among the motivations found in the literature, we highlight the need for transparency (Corrêa et al., 2017), democratization of knowledge (Barcellos et al., 2022), reuse of public data (Rawiro et al., 2023), and the availability and interpretability of information (Barcellos et al., 2022).

To answer research questions RQ1 and RQ2, on identifying processes, bottlenecks, and assessing the quality of public data available on OGD portals, a literature review was conducted to identify exploratory studies and methodologies supporting the understanding, treatment, and standardization of OGD for PM. The search string used was: (“exploratory” OR “analysis” OR “categorization” OR “investigation” OR “application” OR “methodology” OR “process mining tools” OR “open data”) AND (“process mining” AND “government”).

Following structured steps inspired by Rapid Review principles (Cartaxo et al., 2020), studies were selected through mapping, filtering, and data extraction. Searches were performed in Scopus (Elsevier B.V., 2025), for its peer-reviewed content, and Google Scholar (Google LLC, 2025), due to its broad coverage. In Scopus, 45 articles were retrieved. In Google Scholar, with filters (2020–2024, peer-reviewed, excluding citations), 401 results were found. After reviewing titles, abstracts, and keywords, 47 articles mentioning “process mining” and “government data” were selected and fully read to ensure relevance. These studies supported the identification of research gaps and definition of this study’s scope.

The review revealed significant gaps in applying PM to OGD. One major challenge is the lack of methods for processing and selecting OGD for PM tools, especially when contextual information is missing (de Vasconcelos

[1] The data was retrieved from Scopus using the following queries to *egov*, *egov-comp* and *pm*, respectively: (i) TITLE-ABS-KEY((digital AND government) OR (open AND government AND data)); (ii) same as (i) but filtered by subject area “COMP”; (iii) TITLE-ABS-KEY(process AND mining). These strings represent broader research themes connected to this work and were not limited exclusively to OGD. The intention is to illustrate general trends relevant to the study.

et al., 2024a; Rawiro et al., 2023). Although some publications apply PM in public sectors, such as health, administrative workflows, resource optimization, compliance, and citizen services (Chen et al., 2023, 2024; da Costa et al., 2020; de Murillas et al., 2019; Dilmegani, 2024; Drakoulogkonas & Apostolou, 2021; Erdem & Demirörs, 2017; Ito et al., 2020; Rawiro et al., 2023; Unger et al., 2021; Zerbato et al., 2021), there is still a lack of methodologies for categorizing and evaluating data attributes required for PM (de Murillas et al., 2019; Zerbato et al., 2021).

Although PM requires event data, few studies address the transformation of OGD into event logs, an essential step for process discovery (de Murillas et al., 2019). The use of AI techniques to process event logs in the public sector also remains underexplored (Seeliger et al., 2019), and no study was found that comprehensively defines selection criteria or maps key data attributes for PM. Most public data originates from varied sources (databases, ERP systems, transaction logs, spreadsheets, or CSV files der Aalst, 2016) and is typically tabular, requiring transformation into event logs, a complex yet crucial process. Despite these challenges, PM techniques have shown value in the public sector. In Brazil, PM was used to analyze judicial data, improving justice system efficiency (Unger et al., 2021); to identify bottlenecks in regulatory processes within the Federal Executive Branch (da Costa et al., 2020); and to reduce processing times in birth registration (Febrianti, 2024). Still, limitations in infrastructure, trained personnel, and data quality remain critical barriers.

Therefore, some gaps were identified in literature review, like: (i) methods for processing and selecting OGD for PM tools, spatially handling data without context knowledge, (ii) methods for categorizing and evaluating data attributes for process analysis, (iii) studies focused on transforming data into event logs from governmental data, and (iv) the use of artificial intelligence techniques to process existing event logs, adapted to the objectives or needs of governmental organizations. However, in this work we focus is on these gaps (i) and (ii). Given this gaps identified, this study aims to address these challenges by proposing a method for evaluating OGD datasets to process discovery and bottleneck identification, contributing to the broader application of PM techniques in digital government.

Complementary to this, a previous study conducted a comparative analysis of the PM tools most commonly used in academia and the market (de Vasconcelos et al., 2024a). This study, based on a qualitative and quantitative evaluation of tool functionalities and various criteria, served as the foundation for selecting the tools applied here. Building on these findings, an empirically based study was conducted, starting with the choice of the OGD portal, followed by data collection, selection, and evaluation.

Although PM techniques have gained attention in various domains, their application to OGD remains limited, as discussed in Section 2. As identified in the literature review (Section 3), few studies address how to select and assess OGD for use in PM tools, particularly when the data lacks context or is inconsistently structured. These gaps motivated the formulation of this study's Research Questions, as presented in the introduction. The following section presents the methodological approach developed to address them, focusing on the collection, classification, and evaluation of OGD datasets, with particular emphasis on their suitability for process discovery and bottleneck identification.

4. Our Method

For the development of the research, Brazilian OGD was chosen as the source of the case study. The research was organized in steps to collect OGD from the Brazilian open data portal, to evaluate the datasets available on OGD portals: may these datasets can be used to identify processes and bottlenecks? This is our main Research Question (RQ). OGD-PM Evaluation Method (Open Government Data Process Mining Evaluation Method) was divided in steps into collection, selection and evaluation (Figure 1), with the aim of proposing a method for processing OGD datasets, as well as categorizing and selecting the datasets according to the attributes required for PM, covering gaps (i) and (ii). A visual overview of all these stages (Steps 1 to 3) is available online[2].

The choice of Brazilian data stems from the size and diversity of the country, which has a vast population, varied regional interests and numerous public institutions. Despite its potential, many Brazilian cities face limitations that hinder investment in information technology, as well as challenges in the knowledge and application of new technologies for the management and standardized availability of public data (Barcellos et al.,

[2] Visual summary of the OGD-PM Evaluation Method steps, provided in a mind map "Evaluating datasets of open government data for identifying processes".

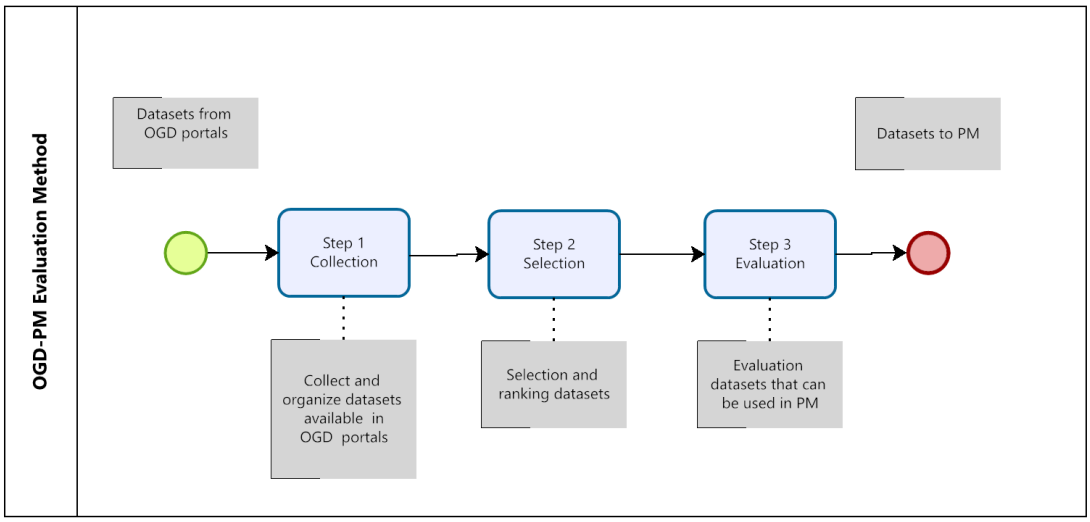

Fig. 1 – OGD-PM Evaluation Method (Open Government Data Process Mining Evaluation Method): Steps of evaluating datasets of open government data for identifying processes and bottlenecks.

2022). This scenario of large scale and diversity makes Brazil an ideal environment for investigating solutions that can improve the transparency and efficiency to population and government institutions. This diversity, combined with the frequent lack of structure in public datasets, makes Brazilian OGD a suitable field for empirical investigation. These characteristics justify its selection for the application of the proposed method, as presented in the Section 5.

4.1. Defining the portal and the sample

To clarify the development of our OGD-PM Evaluation Method (Figure 1), it is important to note that the approach was built empirically, based on exploratory testing and refinement. Initial tests were conducted using publicly available datasets already structured for process mining, allowing the authors to explore initial evaluation parameters and classification strategies. Subsequently, some OGD from ANEEL (the Brazilian Electricity Regulatory Agency) was analyzed[3], selected based on domain familiarity but without knowledge of the internal processes involved. These preliminary experiments are documented in *de Vasconcelos, 2024*. This enabled the identification of challenges typical of OGD, such as data without process context. The method was gradually refined throughout the experimentation conducted on datasets from the Brazilian open data portal.

Therefore, we need to analyze the extent to which the data made available on open data portals can be used in PM tools, to achieve the objective which is possible to apply discovery techniques of PM to OGD, to identify possible bottlenecks and propose improvements. To begin the study, the main Brazilian data repository portals were mapped for potential research. To do this, a search was carried out for articles in the Scopus database (Elsevier B.V., 2025) and the Google Scholar database (Google LLC, 2025) to find publications listing the main portals, in the same manner as the literature review (Section 3). At this point, the aim is to select a portal as the starting point for the Brazilian case study, based on the following search string: “brazilian” AND “open data portals” AND “government”.

The search returned a total of 508 papers, considering the results from 2018 to 2024. The abstracts were analyzed and those that mentioned “Brazilian open data portals” was selected. The research identified that there are independent Brazilian open data portals, but it was identified that dados.gov.br[4] and the Transparency Portal[5] are the main open data portals of the Brazilian Federal Government (Brazil, 2024a, 2024b; Macedo & da Silva Lemos, 2024), published by the Federal Government, various State and Municipal bodies. The Transparency Portal focuses on financial and budgetary data, providing transparency in the use of public resources.

[3] The RALIE dataset (Report on Monitoring the Expansion of Electricity Generation Supply - *Relatório de Acompanhamento da Expansão da Oferta de Geração de Energia Elétrica*) is composed of four open datasets available for download on the ANEEL open data portal here, with an interactive dashboard available here. All materials are in Portuguese.

[4] Open Data Portal (content in Portuguese): https://dados.gov.br/home

[5] Transparency Portal (content in Portuguese): https://portaldatransparencia.gov.br/

The dados.gov.br portal brings together a variety of datasets with diverse content, promoting the reuse of data for different purposes. Given the range of information and the aim of the research, dados.gov.br was chosen as the data source for this study.

Based on this, the information available on the Brazilian Federal Government's Open Data Portal - dados.gov.br (Brazil, 2024a) was analyzed, which provided 12,789 datasets verified on 14 March 2024. For statistical analysis of the data, a systematic representative sample of the population was carried out, with size calculated based in equation Eq.1, where n represents the selected sample size, N the population size referring to the dataset available on the portal, Z the standardized z-score value corresponding to a 95% confidence level (which in this case was 1.96), p the standard deviation of 50% to ensure the sample is sufficiently large, and e the margin of error set at 0.05, determining the maximum permissible error. Based on these parameters, the calculated sample size was 373 datasets to be collected from the portal.

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1-p)}{Z^2 \cdot p \cdot (1-p) + e^2 \cdot (N-1)} \tag{1}$$

4.2. Cataloging the data obtained

The collection stage primarily involves defining the OGD portal, the sample, and the organization of the datasets, involving steps of mapping the OGD portals, defining the chosen portal, defining the sample, collecting the datasets, organizing the datasets and classifying them according to relevance, prioritizing event logs and tables with temporal data, ensuring a structured collection in line with the study's objectives. The data was organized with metadata information obtained from the OGD portal such as the "originating agency", "links" and "file names". As an important part of the proposed method, another metadata defined was the "categorization", established on the type of data collected.

Therefore, the datasets has been organized and categorized for the analysis and selection of data that are relevant to the objectives of this research. This information were created from the data collected to standardize information and not from a pattern defined on the portal. The description as shown in Table 1, according to the datasets available on the dados.gov.br open data portal: assimilating the title (i), the body responsible for the data (ii), the download link (iii), the name of the file (iv) and its categorization (v) in order to identify if the file refers to an event log (1), if there is a complete date that can be used as a Timestamp (2), if it consists a list of data or an information table (3), documents (4), if there is no file at all (5) or just a link to another portal (6). The full list of categorizing datasets can be found in online repository (de Vasconcelos et al., 2024c) of datasets and more information.

Tab. 1 – Description of the categorized datasets obtained from the Brazilian Federal Government's Open Data Portal - dados.gov.br[4]

Column	Name	Description	Domain of collected data
(i)	Dataset	Title of the dataset	Text
(ii)	Organization	Originating organization	Text
(iii)	Link	Link to the data on dados.gov.br	Text
(iv)	File	Name of the file	Text
(v)	Categorization	Categorization of the obtained dataset	(1) Event logs
			(2) Data table with date
			(3) Data table (list)
			(4) Documents
			(5) Nonexistent file
			(6) Nonexistent file - external link

Based on the categories defined in item (v), Table 1, in order to verify whether the datasets could be used in PM tools, the data obtained from the dados.gov.br portal was analyzed and classified as "Relevant" or "Not Relevant" for this research. Among these, only categories (1) and (2) were considered "Relevant" for this research, as they met the necessary conditions for process mining. The remaining categories were classified as "Not Relevant" due to lack of required structure or accessibility. The complete table with all the datasets and their relevance classification is available at the online repository[6] (de Vasconcelos et al., 2024b).

[6] Link for the complete table with the results of the analysis the data obtained from the dados.gov.br portal, classifying them as "Relevant" and "Not Relevant" for the research: open_data_sample.xlsx

Although categories (3), (4) and (6) could be more deeply analyzed and methods for transforming and adapting the data studied, we didn't go into this merit in the study, opting for more objective resources for the preliminary development of the research, choosing categories (1) and (2) as relevant. These decisions are grounded in the core requirements for PM, which include the presence of identifiable case identifiers, timestamps, and activities. Only datasets categorized as (1) Event logs and (2) Data tables with date were found to potentially satisfy these conditions, justifying their classification as "Relevant".

4.3. Definition of ranking criteria

In order to make a ranking of the selected datasets to classify them according to their score, criteria to assessment metadata from datasets were created according to the information (In), importance (I), difficulty (D) and ease (E) of manipulation. These criteria were developed empirically, based on the practical challenges faced during the development of the research, based on tests carried out with various datasets. In this way, the information contained in each dataset was qualified according to these criteria. The focus was to create criteria that reflect specific demands of process mining applications, and their structure was refined iteratively as the analysis progressed. Table 2 shows all the evaluation criteria and their domains used to carry out the study. In items 6, 7 and 8, the integer domain from 1 to 5 corresponds to 1 minimum and 5 maximum values. The criteria were refined iteratively as the study progressed and were specifically designed to support the identification of datasets most suitable for PM. Further information on this research can be found in an online repository (de Vasconcelos et al., 2024b).

Tab. 2 – Evaluation and qualification criteria for the selected datasets

Item #	In/I/D/E	Criteria	Domain
1	In	Number of records	Integer
2	In	Number of columns	Integer
3	In	Number of columns with dates	Integer
4	I	Part of the 17 Goals	Yes or No (1 or 0)
5	I	Public Policies prioritized by the PPA	Yes or No (1 or 0)
6	D	Difficulty based on the number of records	Integer from 1 to 5
7	D	Difficulty based on the number of columns	Integer from 1 to 5
8	D	Difficulty in identifying processes	Integer from 1 to 5
9	D	Records with perceptible issues	Yes or No (1 or 0)
10	D	Need to transform the data (columns to rows)	Yes or No (1 or 0)
11	E	Event log	Yes or No (1 or 0)
12	E	Data standardization	Yes or No (1 or 0)

The information criteria (In), which deal with the number of records (rows/instances), number of columns (features) and columns with dates (items 1, 2 and 3), received their nominal values identified in each dataset. Importance (I) was assessed on the basis of its significance for public policies (items 4 and 5), difficulty (D) on the basis of the complexity of manipulating the data (items 6 to 10) and ease (E) on the simplicity of using the data in mining tools (items 11 and 12). The importance of the data, as it is a subjective criterion, was measured based on belonging or not (Yes or No) to the 17 UN Sustainable Development Goals (ONU, 2015) and the public administration's priority public policies in the Multi-Year Plan (PPA) (Brasil et al., 2023). The numbered list of priorities can be found in the online repository (de Vasconcelos et al., 2024b).

The difficulty related to the number of records and columns (items 6 and 7) was calculated using the raw values previously stored as informational criteria (items 1 and 2). These values were sorted in ascending order, and a frequency distribution analysis was performed. Based on this, the values were grouped into five classes using a class interval method, which divides the full range of values into equal-width intervals (Montgomery & Runger, 2018). This approach facilitates transforming continuous variables into discrete difficulty levels. The difficulty rating for items 6, 7 and 8 ranged from 1 (least difficulty) to 5 (greatest difficulty). For the number of records, the classes were defined as follows: Class 1 for 1 to 300 records, Class 2 for 301 to 2,000, Class 3 for 2,001 to 4,000, Class 4 for 4,001 to 100,000, and Class 5 for more than 100,000 records. For the number of columns, Class 1 included datasets with 6 to 13 columns, Class 2 with 14 to 20, Class 3 with 21 to 27, Class 4 with 28 to 34, and Class 5 with 35 to 38 columns.

The criteria that identify the presence of noticeable problems in the records (item 9), the need to transform data from columns to rows (item 10), whether the file is an event log (item 11) and the existence of standard-

ization in the data (item 12) were evaluated individually, each receiving a "Yes" or "No" answer as appropriate. In order to properly rank the datasets, the results were normalized to standardize the results, the calculations used for normalization are described in Appendix A.

5. Analysis of dados.gov.br

The Brazilian OGD Portal – dados.gov.br(Brazil, 2024a), launched in 2012, is the central platform for accessing public data published by Brazilian government agencies. Each organization is responsible for its own data. The portal currently hosts around 14k datasets across diverse areas such as supply, administration, agriculture, fishing, communications, and national defense. Due to its volume, heterogeneity, and lack of standardization, dados.gov.br was chosen as the case study for this research, representing a realistic scenario to test the applicability of PM techniques to OGD in complex and large-scale environments. Based on the sampling method described in the methodology, 373 datasets were collected by selecting every 20th item from the portal's unordered list. All collected datasets are available in an online repository (de Vasconcelos et al., 2024c).

Based on the categories defined in Table 1, the datasets from dados.gov.br were analyzed and classified by their relevance to PM tools. The analysis presented practical challenges: each file had to be manually examined to identify potential processes, which was particularly time-consuming when dealing with large datasets with many instances and features (columns and records). Long files made it harder to detect distinct activities, requiring an iterative approach to narrow the focus. Additionally, the lack of standardization across datasets demanded constant transformations, increasing the workload in this phase.

As a result, most of the data was found unsuitable for PM tools, often due to corrupted files, empty links, or redirections to other portals. Some datasets consisted only of documents (e.g., PDF, DOC) or simple lists without timestamps. Although redirected portals may offer usable data, analyzing external sources was beyond the scope of this study. The proposed method focuses on evaluating data directly from the main OGD platform. Consequently, 24% of the sample (89 datasets) was deemed relevant, 23% in dated table format and 1% as event logs. The remainder was excluded. Details of this methodology are in Section 4.2, and the full relevance classification is available in de Vasconcelos et al., 2024b.

5.1. Preliminary selection of datasets

In the previous step, 89 datasets were identified as relevant to this research. However, analyzing all of them in depth would require considerable time and effort. Since this study represents an exploratory effort to demonstrate the feasibility and applicability of the proposed evaluation method, we opted to apply the method to a representative subset of this data. To select this subset, we used the Pareto principle (Koch, 1999), which suggests that a small portion of data can often yield most of the valuable insights[7]. We then selected 24 datasets for this analysis (just over 20% of the data considered relevant) as follows: all the "Event Logs" identified in the sample (4 datasets) and 20 categorized as "Data table with date". The selected datasets are listed in Table 3. The titles have been translated but the originals are in Portuguese.

Of the sets categorized as "Data table with date", the topic, update date (from 2020 to 2024) and origin of the data were taken into account, so that a maximum of three data sets from the same government agency were selected to ensure data variability. We then selected data from 16 different government organizations, such as: ANAC, ANATEL, ANEEL, ANVISA, BCB, CGU, FURG, IBAMA, IFBA, IFCE, INPI, MGI, PBH, TSE, UFG and UFRB.

Therefore, the 24 selected datasets were organized, analyzed individually according to each criterion defined in this section and the information obtained was recorded, as shown in Table 4. Table 2 shows all the evaluation criteria and their domains used to carry out the study. As part of the difficulty criterion (D), item 8, "Difficulty in identifying processes", the data from each dataset was analyzed in order to identify which process should be analyzed first. In addition, the key attributes were identified (mapping of "caseid", "timestamps" and "activities") which served as the starting point for each dataset. The time spent on this analysis was recorded. Finally, the average time for the initial analysis and mapping was 77 minutes, used to calculate the total analysis time.

[7] The Pareto Principle (Koch, 1999), also known as the 80/20 rule, states that 80% of relevant information can be obtained from 20% of the available data

Tab. 3 – Datasets selected for preliminary analysis

ID	Dataset	Agency	Update	Cat.
002	R&D Projects in Electric Energy	ANEEL	02/01/2024	2
003	SIGA - ANEEL Generation Information System	ANEEL	02/01/2024	2
040	Last Day Contracts	MGI	02/10/2024	2
061	Aeronautical Safety Recommendations	ANAC	02/11/2024	2
096	CNEP - National Registry of Punished Companies	CGU	02/10/2024	2
128	COVID-19 Spreadsheet 07-27-2020	PBH	07/28/2020	2
133	Industrial Design Registration Requests (2018)	INPI	01/14/2021	2
137	Computer Program Registration Requests (2020)	INPI	02/16/2021	2
138	Public Commitments Agenda of the Director of Patents, Computer Programs, and Integrated Circuit Topographies at INPI (2017)	INPI	02/17/2021	2
166	Aircraft - Registered Drones	ANAC	09/09/2021	2
186	Money in Circulation - Detailed Daily Information	BCB	02/13/2024	2
195	Offered Courses	IFBA	05/17/2022	2
205	Spectrum and Orbit - Satellites	ANATEL	06/10/2022	2
206	Competition - Infrastructure Sharing Contracts	ANATEL	10/23/2020	2
247	Periodicals	UFRB	10/19/2022	1
251	Results - 2016 - Expected and Actual Correspondences - 1st Round	TSE	10/25/2022	1
253	Extension Actions	IFCE	10/26/2022	2
258	Registered Diplomas	UFG	11/08/2022	2
270	Post-Market Hemovigilance	ANVISA	01/12/2023	2
284	Results - 2008	TSE	03/04/2023	1
307	Fala.BR - Access to Information Module	CGU	06/14/2023	2
316	Sinaflor - POA Amazon Legal	Ibama	07/08/2023	2
356	FURG - Registration of Fiscal Officers for Public Tenders	FURG	10/02/2023	1
357	ANEEL Consumer Satisfaction Index (IASC)	ANEEL	11/22/2023	2

Tab. 4 – Result of comparative analysis of datasets based on Informative Data (In), Importance (I), Difficulty (D), and Ease of Use (E). Table 2 shows details relating number 1 until 12 evaluation criteria.

#	Informative Data (In)			Importance (I)		Difficulty (D)					Ease of use (E)	
Dataset	1	2	3	4	5	6	7	8	9	10	11	12
002	3718	17	3	Yes	Yes	3	2	5	Yes	No	No	Yes
003	25909	23	4	Yes	Yes	4	3	1	Yes	Yes	No	Yes
002	3718	17	3	Yes	Yes	3	2	5	Yes	No	No	Yes
003	25909	23	4	Yes	Yes	4	3	1	Yes	Yes	No	Yes
040	1	34	4	No	No	1	4	5	Yes	Yes	No	No
061	3090	14	4	No	Yes	3	2	1	Yes	Yes	No	Yes
096	906	21	4	Yes	No	2	3	1	Yes	Yes	No	Yes
128	130	29	3	Yes	Yes	1	4	2	Yes	Yes	No	No
133	80161	21	6	Yes	No	4	3	1	Yes	No	No	No
137	3122	17	4	Yes	No	3	2	1	No	No	No	Yes
138	213	12	1	No	No	1	1	5	Yes	Yes	No	No
166	135287	10	1	No	No	5	1	5	Yes	Yes	No	No
186	198195	6	1	No	No	5	1	5	No	Yes	No	Yes
195	50525	23	2	Yes	Yes	4	3	5	Yes	Yes	No	Yes
205	723	26	3	No	No	2	3	1	No	No	No	Yes
206	8207	11	3	Yes	Yes	4	1	5	Yes	Yes	No	No
247	12	13	1	Yes	Yes	1	1	5	Yes	Yes	Yes	No
251	1780	13	2	No	No	2	1	5	Yes	No	Yes	Yes
253	2104	11	2	Yes	Yes	3	1	3	No	Yes	No	Yes
258	79	15	5	Yes	Yes	1	2	1	No	Yes	No	Yes
270	189403	16	2	Yes	Yes	5	2	2	No	Yes	No	Yes
284	646942	38	2	No	No	5	5	3	No	No	Yes	Yes
307	1756	18	3	No	No	2	2	3	No	Yes	No	Yes
316	57573	50	5	Yes	Yes	4	5	5	Yes	No	No	Yes
356	2517	6	2	No	Yes	3	1	2	Yes	No	Yes	Yes
357	1185	159	1	No	No	2	5	5	Yes	Yes	No	Yes

Although no specific weight was applied to the criteria, it is recognized that an event log or certain information, such as “caseid”, “timestamps” and “activities” identified, are crucial to the study of the data. In addition, the data was selected from an initial sample of 373 datasets and subjected to various filtering steps in order to ensure a choice of the most up-to-date and relevant data possible. The normalization was calculated as shown in Appendix A.

5.2. Ranking of datasets

With the computed values, it was possible to identify that 52% of the selected datasets were classified as important for society because they are part of public policies, 55% present a degree of difficulty for manipulation in the mining tools and 44% show no apparent difficulty for use in the PM tools. Based on the criteria presented, the final classification of the data sets was reached. It was possible to identify that the 80th percentile is less than 6 (six), i.e. 80% of the data achieved a score of less than 6 (six) and 50% of the cases achieved a score of less than 5 (five). The normalized values and complete ranking data can be found in online repository (de Vasconcelos et al., 2024b). The result of ranking by normalized criteria (Appendix A) of selected datasets are described in Table 5.

Tab. 5 – Result of ranking by normalized criteria of selected datasets based on Informative Data (In), Importance (I), Difficulty (D), and Ease of Use (E). Table 2 shows details relating number 3 until 12 evaluation criteria.

#	Inf. (In)	Importance (I)		Difficulty (D)					Ease of use (E)		Classification	
Dataset	3	4	5	6	7	8	9	10	11	12	Score	Rank
258	1.00	1.00	1.00	1.00	0.75	1.00	1.00	0.00	0.00	1.00	7.75	1st
137	0.80	1.00	0.00	0.50	0.75	1.00	1.00	1.00	0.00	1.00	7.05	2nd
356	0.40	0.00	1.00	0.50	1.00	0.75	0.00	1.00	1.00	1.00	6.65	3rd
253	0.40	1.00	1.00	0.50	1.00	0.50	1.00	0.00	0.00	1.00	6.40	4th
270	0.40	1.00	1.00	0.00	0.75	0.75	1.00	0.00	0.00	1.00	5.90	5th
002	0.60	1.00	1.00	0.50	0.75	0.00	0.00	1.00	0.00	1.00	5.85	6th
205	0.60	0.00	0.00	0.75	0.50	1.00	1.00	1.00	0.00	1.00	5.85	7th
003	0.80	1.00	1.00	0.25	0.50	1.00	0.00	0.00	0.00	1.00	5.55	8th
316	1.00	1.00	1.00	0.25	0.00	0.00	0.00	1.00	0.00	1.00	5.25	9th
247	0.20	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	5.20	10th
061	0.80	0.00	1.00	0.50	0.75	1.00	0.00	0.00	0.00	1.00	5.05	11th
096	0.80	1.00	0.00	0.75	0.50	1.00	0.00	0.00	0.00	1.00	5.05	12th
284	0.40	0.00	0.00	0.00	0.00	0.50	1.00	1.00	1.00	1.00	4.90	13th
133	1.00	1.00	0.00	0.25	0.50	1.00	0.00	1.00	0.00	0.00	4.75	14th
128	0.60	1.00	1.00	1.00	0.25	0.75	0.00	0.00	0.00	0.00	4.60	15th
307	0.60	0.00	0.00	0.75	0.75	0.50	1.00	0.00	0.00	1.00	4.60	16th
195	0.40	1.00	1.00	0.25	0.50	0.00	0.00	0.00	0.00	1.00	4.15	17th
251	0.40	0.00	0.00	0.75	1.00	0.00	0.00	0.00	1.00	1.00	4.15	18th
206	0.60	1.00	1.00	0.25	1.00	0.00	0.00	0.00	0.00	0.00	3.85	19th
186	0.20	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	3.20	20th
138	0.20	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	2.20	21st
040	0.80	0.00	0.00	1.00	0.25	0.00	0.00	0.00	0.00	0.00	2.05	22nd
357	0.20	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	1.00	1.95	23rd
166	0.20	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.20	24th

With this in mind, the three best-scoring datasets were selected for testing in the tools, namely: (i) 258 - Registered Diplomas, (ii) 137 - Computer Program Registration Requests (2020) and 356 - FURG - Registration of civil servants for Public Tenders, two from universities (UFG and FURG) and one from the INPI. 258 requires columns to be transformed into rows, the others don’t, and only 356 is an event log. The other event logs did not score well due to various problems found in the data. Although they are event logs, some did not have complete information, making it difficult to define a process to be analyzed (characterized only as a poorly identified event log with a lack of documentation). Another problem was the excess of instances (records), making these the main reasons for the low score of such datasets.

5.3. Evaluating the datasets and preliminary tests

Given the nature of OGD, which often lacks explicit processes, a preliminary evaluation was conducted to identify datasets with attributes (features) essential for process discovery and bottleneck identification. Al-

though the evaluation began early in the research, further analysis was needed before applying datasets to PM tools—such as verifying the presence of minimum required attributes and assessing whether the processes discovered made sense for each dataset. As this study focuses on evaluating datasets for process discovery, these initial tests helped confirm the potential use of the selected datasets.

To guide this assessment, datasets were examined based on their structural adequacy for PM, and availability of key attributes such as event identifiers, timestamps, and activity descriptions. These attributes play a fundamental role in enabling the reconstruction of process flows and the identification of inefficiencies. As a result of this evaluation, a subset of datasets was selected for preliminary testing using the PM tools Disco, ProM and Celonis, chosen to carry out the tests. In order to choose the most suitable PM tools for comparing the results proposed in this study, a previous study was used to evaluate the PM tools used in academia and the market (de Vasconcelos et al., 2024a). This study evaluated the tools using qualitative and quantitative criteria and ranked them, among other criteria, based on user adherence, relevance to the market and academia and objectivity in generating results. Another study that also served as a reference carried out an analysis, based on a structured methodology that allows the analyst to compare any number of PM tools using any number of comparative analysis criteria (Drakoulogkonas & Apostolou, 2021).

As for the content of the first dataset selected for testing, dataset 258, the results indicate that the data is clean, i.e. there are no records of bottlenecks or process deviations, only data that has been successfully completed. In addition, a comparative analysis between the tools shows that there is no significant difference in the tests, with the variation in average time between activities showing little difference between the discovery algorithms of each tool. In terms of use, it was found that loading data into the tools was considerably faster in the Disco and ProM tools. Dataset 137 was more complex, requiring re-analysis of the mapping; however, the process graph was successfully generated, and the information was clear. It was found that the process occurs relatively quickly, but with some deviations that increase the completion of cases. As for the comparative analysis of the content, the results showed a significant difference when calculating the average duration of the cases. As for the usage analysis, the data loading time varied greatly and required advanced analysis in the ProM tool. Figure 2 shows the process discovered and its bottlenecks in the three tools tested.

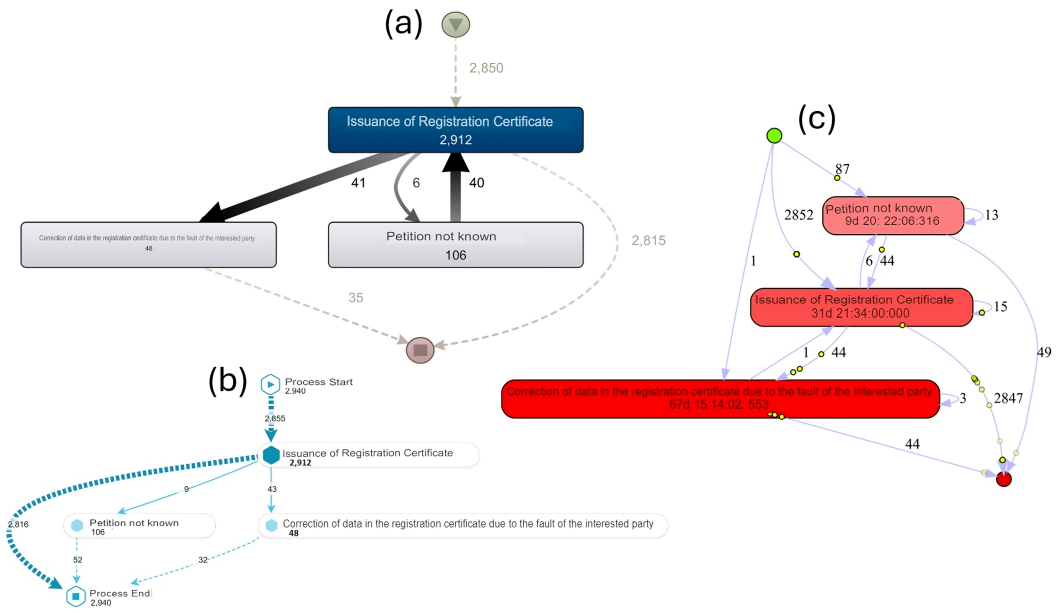

Fig. 2 – Process map discovered with dataset 137 - Computer Program Registration Requests (2020) in (a) in Disco, (b) Celonis and (c) ProM. These process maps highlight the issuance of registration certificates as the main activity, with variations in visualization and process details across tools. ProM emphasizes execution times, revealing a bottleneck in data correction (67 days), while Disco and Celonis focus on process flows and frequencies.

In the last experiment, dataset 356, the level of complexity increased considerably, marked by difficulties in understanding the data, lack of standardization, and re-analysis of the mapping. Analyzing the content, the data was only similar when an advanced analysis was carried out on the tools, and even then, with differences in the main activities and the longest average time. As for the use of the tools, ProM was the one that responded best to the complexity of this dataset. From these experiments, it was possible to see the difficulty in preparing this data, made available on open data portals, for testing in the PM tools. This is evident from the lack of reference documentation (hindering understanding) and the increased complexity given the amount of data found in the datasets (quantities of records and columns to be analyzed). Although challenging, the difficulty in pre-processing the data can be measured by ranking the datasets. In addition, despite the difficulties mentioned, it was possible to analyze the data, establish processes, identify bottlenecks, points of attention in the datasets and recommendations.

6. Conclusions and Future Work

This study addressed the challenges in evaluating OGD for PM applications, focusing on the collection, selection, and assessment of datasets. The findings highlight the difficulties in handling datasets without predefined context, a common characteristic of OGD, which requires structured evaluation criteria to determine their relevance and suitability for process discovery. It was possible to identify datasets with the necessary attributes for PM, demonstrating that, despite inconsistencies in data representation and quality, these datasets can be adapted for meaningful analysis. The study reinforces the importance of well-defined methodologies for assessing data structure, selecting relevant datasets, and verifying their adequacy for PM, ensuring that public data can be effectively used to understand and improve government processes.

This work contributes to the literature by addressing two gaps identified in the field: the lack of (i) methods for processing and selecting OGD for PM tools, especially when dealing with data without context knowledge, and also (ii) of methods for categorizing and evaluating data attributes to process analysis. The OGD-PM Evaluation Method responds to these gaps by offering a transparent and replicable procedure for assessing OGD datasets. While many existing studies apply PM to predefined event logs or within specific domains, this work offers a generalizable approach that can be adapted to varied OGD, including those without structured metadata or contextual information. This flexibility makes the method particularly valuable for advancing future research on OGD, supporting broader adoption of PM in digital government contexts.

The findings of this study provide a direct response to the proposed research questions. Regarding RQ1, the results demonstrate that OGD can be evaluated and applied in PM tools to discover processes and identify bottlenecks in public services. Despite the challenges of data inconsistencies and the lack of contextual metadata, the approach used in dataset selection and evaluation made it possible to identify relevant datasets and successfully apply them to PM analyses. In response to RQ2, the study established key criteria for selecting and preparing OGD, emphasizing the need for evaluation based on data suitability for PM. The research reveals that even in the absence of a predefined context, it is possible to extract meaningful insights from OGD when applying PM. These contributions reinforce the importance of methodologies for OGD analysis, ensuring its effective use in understanding and optimizing government processes.

Our work contributes to PM practices on OGD, providing steps for the evaluation and preparation of datasets that facilitate the identification of processes and operational bottlenecks. From this work, researchers and professionals can evaluate and prepare OGD for PM analyses, to identify the inherent processes within the data and detect bottlenecks, thereby enhancing the understanding of the data and the efficiency of governmental processes. The limitations of this research are primarily associated with the generalization of the results due to the diversity of the OGD and the categorization method applied, which may not capture all the nuances of the varied datasets. Moreover, the results may vary according to the quality and specific structure of the available data. Future research could extend the application of the proposed method to the full set of relevant datasets, allowing for broader validation and refinement of the evaluation framework. Additional studies may also explore datasets currently classified as not relevant, to assess the boundaries of their applicability to PM and further test the evaluation criteria. Expanding the method to include datasets accessed via external links and developing automated tools to support large-scale assessment and selection are also promising directions. These extensions may support more analyses, promoting a deeper understanding of how PM can be applied to improve governmental processes.

Acknowledgement

The authors would like to thank the Graduate Program in Computer Science at Fluminense Federal University (PGC/UFF) for the institutional support provided throughout this research, which were essential for the development and dissemination of this work.

Funding or Grant

This work was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under the Public Call No. 23/2018 – Academic Doctorate for Innovation Program (*Doutorado Acadêmico para Inovação* – DAI), grant number 161128/2019-0.

Data/Software Access Statement

All datasets used in this study are publicly available and were obtained from the Brazilian Federal Government's Open Data Portal (dados.gov.br). The full list of datasets analyzed, along with metadata, classification details, and the evaluation criteria used in the study, is available in the following repositories: (i) https://doi.org/10.6084/m9.figshare.25514884.v5 and (ii) https://doi.org/10.17632/x8sgcykthn.1.

Contributor Statement

Author 1 contributed to Conceptualization, Methodology, Data Curation, Investigation, Formal Analysis, Visualization, and Writing – Original Draft. Author 2 and Author 3 contributed to Conceptualization, Supervision, and Writing – Review & Editing.

Conflict Of Interest (COI)

There is no conflict of interest.

References

Barcellos, R., Bernardini, F., & Viterbo, J. (2022). Towards defining data interpretability in open data portals: Challenges and research opportunities. *Information Systems*, *106*. DOI: https://doi.org/10.1016/j.is.2021.101961.

Barcellos, R., Bernardini, F., Zuiderwijk, A., & Viterbo, J. (2024). Exploring interpretability in open government data with chatgpt. *Proceedings of the 25th Annual International Conference on Digital Government Research*, 186–195. DOI: https://doi.org/10.1145/3657054.3657079.

Berners-Lee, T. (2006, July). Linked data - design issues. https://www.w3.org/DesignIssues/LinkedData

Brasil, do Planejamento e Orçamento, M., & de Planejamento, S. N. (2023). Plano plurianual 2024-2027: Mensagem presidencial/ministério do planejamento e orçamento, secretaria nacional de planejamento. https://www.gov.br/planejamento/pt-br/assuntos/plano-plurianual/paginas/lei-do-ppa-2024-2027%20https://www.gov.br/planejamento/pt-br/assuntos/plano-plurianual/arquivos/projeto-de-lei-ppa-2024-2027/projeto-de-lei-ppa2024-2027.pdf

Brazil. (2024a). Brazilian federal government's open data portal - portal brasileiro de dados abertos. https://dados.gov.br/home

Brazil. (2024b). What it is and how it works. federal government transparency portal. https://portaldatransparencia.gov.br/sobre/o-que-e-e-como-funciona

Burke, A. T., Leemans, S. J., Wynn, M. T., van der Aalst, W. M., & ter Hofstede, A. H. (2024). A chance for models to show their quality: Stochastic process model-log dimensions. *Information Systems*, *124*, 102382. DOI: https://doi.org/10.1016/J.IS.2024.102382.

Cartaxo, B., Pinto, G., & Soares, S. (2020). Rapid reviews in software engineering. In M. Felderer & G. H. Travassos (Eds.), *Contemporary empirical methods in software engineering* (pp. 357–384). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-32489-6_13.

Chen, K., Abtahi, F., Carrero, J.-J., Fernandez-Llatas, C., & Seoane, F. (2023). Process mining and data mining applications in the domain of chronic diseases: A systematic review. *Artificial Intelligence in Medicine*, *144*, 102645. DOI: https://doi.org/https://doi.org/10.1016/j.artmed.2023.102645.

Chen, K., Abtahi, F., Carrero, J.-J., Fernandez-Llatas, C., Xu, H., & Seoane, F. (2024). Proposing a novel methodology for process mining in clinical epidemiology: Insights and validation from a case study on chronic kidney disease progression. *SSRN*.

Corrêa, A. S., de Paula, E. C., Corrêa, P. L. P., & da Silva, F. S. C. (2017). Transparency and open government data: A wide national assessment of data openness in brazilian local governments. *Transforming Government: People, Process and Policy*, *11*, 58–78. DOI: https://doi.org/10.1108/TG-12-2015-0052/FULL/PDF.

da Costa, H. P., de Assis, J. M. V., & de Vasconcelos, C. C. (2020, January). Case study: Government process mining in the brazilian executive branch. https://www.tf-pm.org/resources/casestudy/government-process-mining-in-the-brazilian-executive-branch

de Murillas, E. G. L., Reijers, H. A., & van der Aalst, W. M. (2019). Connecting databases with process mining: A meta model and toolset. *Software and Systems Modeling*, *18*, 1209–1247. DOI: https://doi.org/10.1007/S10270-018-0664-7/FIGURES/34.

de Oliveira, D., de Oliveira, D. G., & Filho, O. O. (2024). Dados abertos da previdência social: Um estudo avaliativo. *Revista Meta: Avaliação*, *0*. DOI: https://doi.org/10.22347/2175-2753v0i0.4797.

de Vasconcelos, G. S. (2024, October). *Process mining in open government data* [PhD thesis]. Fluminense Federal University, Computer Science Institute. https://sucupira-legado.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=15574584

de Vasconcelos, G. S., Bernardini, F., & Viterbo, J. (2024a). A comparison between the most used process mining tools in the market and in academia: Identifying the main features based on a qualitative analysis. In A. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & V. Colla (Eds.), *Information systems and technologies* (pp. 218–228). Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-45645-9_21.

de Vasconcelos, G. S., Bernardini, F., & Viterbo, J. (2024b). Criteria for evaluating and qualifying public datasets obtained from the brazilian federal government's open data portal - dados.gov. DOI: https://doi.org/10.17632/X8SGCYKTHN.1.

de Vasconcelos, G. S., Bernardini, F., & Viterbo, J. (2024c, March). Datasets obtained from the brazilian federal government's open data portal - dados.gov. DOI: https://doi.org/10.6084/m9.figshare.25514884.v5.

der Aalst, W. V. (2016). *Process mining: Data science in action*. Springer. DOI: https://doi.org/10.1007/978-3-662-49851-4.

Dilmegani, C. (2024). Top 44 process mining use cases applications in 2024. *AIMultiple Research*. https://research.aimultiple.com/process-mining-use-cases/

Drakoulogkonas, P., & Apostolou, D. (2021). On the selection of process mining tools. *Electronics (Switzerland)*, *10*, 1–24. DOI: https://doi.org/10.3390/electronics10040451.

Elkoumy, G., Pankova, A., & Dumas, M. (2023). Differentially private release of event logs for process mining. *Information Systems*, *115*. DOI: https://doi.org/10.1016/j.is.2022.102161.

Elsevier B.V. (2025). Scopus preview - scopus - welcome to scopus. https://www.scopus.com/home.uri

Erdem, S., & Demirörs, O. (2017). An exploratory study on usage of process mining in agile software development. *Communications in Computer and Information Science*, *770*, 187–196. DOI: https://doi.org/10.1007/978-3-319-67383-7_14.

Febrianti, N. A. (2024). Analisis proses pengajuan akta kelahiran dispendukcapil surabaya menggunakan process mining untuk mempercepat waktu proses. *ITS*. https://repository.its.ac.id/105592/

Google LLC. (2024). Google trends: Search term "process mining", from 2004 to 2024, in brazil. https://trends.google.com/trends/explore?date=all&geo=BR&q=process%20mining&hl=pt-BR

Google LLC. (2025). About google scholar. https://scholar.google.com.br/intl/pt-BR/scholar/about.html

Heumann, M., Kraschewski, T., Werth, O., & Breitner, M. H. (2024). Reassessing taxonomy-based data clustering: Unveiling insights and guidelines for application. *SSRN*. DOI: https://doi.org/10.2139/SSRN.4716206.

Ito, S., Vymětal, D., & Šperka, R. (2020). Process mining approach to formal business process modelling and verification: A case study. *Journal of Modelling in Management*, *16*. DOI: https://doi.org/10.1108/JM2-03-2020-0077.

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2019). The sustainable value of open government data. *Journal of the Association for Information Systems*, *20*, 6. DOI: https://doi.org/10.17705/1jais.00549.

Kerremans, M., Iijima, K., Sachelarescu, A. R., Duffy, N., & Sugden, D. (2023, March). Magic quadrant for process mining tools. https://www.gartner.com/en/documents/4192799

Kerremans, M., Sugden, D., & Duffy, N. (2024, April). Magic quadrant for process mining platforms. https://www.gartner.com/doc/reprints?id=1-2HGMM7VN&ct=240502&st=sb

Koch, R. (1999). *The 80/20 principle: The secret of achieving more with less* (3a). Crown Business.
Lanoue, J. (2020). Disparate environmental monitoring as a barrier to the availability and accessibility of open access data on the tidal thames. *Publications*, *8*(1), 6.
Macedo, D. F., & da Silva Lemos, D. L. (2024). Open government data: Maturity diagnosis model for quality data published on the web. *Em Questão*, *30*, 1–14. DOI: https://doi.org/10.1590/1808-5245.30.132617.
Montgomery, D. C., & Runger, G. C. (2018). *Applied statistics and probability for engineers* (7th). Wiley.
ONU. (2015). Objetivos de desenvolvimento sustentável. https://brasil.un.org/pt-br/sdgs
Parks, W. (1957). The open government principle: Applying the right to know under the constitution. *The George Washington Law Review*, *26*, 1–22. https://heinonline.org/HOL/LandingPage?handle=hein.journals/gwlr26&div=10&id=&page=
Rajabi, E., Midha, R., & de Souza, J. F. (2023). Constructing a knowledge graph for open government data: The case of nova scotia disease datasets. *Journal of Biomedical Semantics*, *14*, 1–10. DOI: https://doi.org/10.1186/S13326-023-00284-W/FIGURES/5.
Rawiro, D., Gaol, F. L., Supangkat, S., & Ranti, B. (2023). Process mining applications in government sector: A systematic literature review. *IEOM*, 1–14. DOI: https://doi.org/10.46254/eu05.20220617.
Seeliger, A., Guinea, A. S., Nolle, T., & Mühlhäuser, M. (2019). Processexplorer: Intelligent process mining guidance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11675 LNCS*, 1]. DOI: https://doi.org/10.1007/978-3-030-26619-6_15.
Unger, A. J., Neto, J. F. D. S., Fantinato, M., Peres, S. M., Trecenti, J., & Hirota, R. (2021). Process mining-enabled jurimetrics: Analysis of a brazilian court's judicial performance in the business law processing. *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*, 240–244. DOI: https://doi.org/10.1145/3462757.3466137.
Vasconcelos, L., Barcellos, R., Viterbo, J., Bernardini, F., Salgado, L., & Trevisan, D. (2020). Investigating communicability issues in the open data manipulation flow. *AMCIS 2020 Proceedings*. https://aisel.aisnet.org/amcis2020/sig_hci/sig_hci/18
Zerbato, F., Soffer, P., & Weber, B. (2021). Initial insights into exploratory process mining practices. *Lecture Notes in Business Information Processing*, *427 LNBIP*, 145–161. DOI: https://doi.org/10.1007/978-3-030-85440-9_9.

Appendix A

Calculations for normalization

As each criterion for the selected datasets (Table 2) had its own way of scoring due to its characteristics, in order to make it easier to compare them and standardize them, it was decided to normalize the scores on a scale of 0 to 1. The overall score, which enabled the final classification of the data sets, ranged from 0 to 10, where 0 (zero) corresponds to a minimum value and 10 to a maximum value. For the criteria that assess difficulty, the score was 0 (zero) for the greatest difficulty, and 1 for minimum difficulty. Criteria 1 and 2 were used to generate 6 and 7 respectively, as seen in the previous section, so normalization was applied from criteria 3 to 12. Calculations for normalization were based on the following conversions, described in the table 6 and details can be found at de Vasconcelos et al. (2024b).

Tab. 6 – Normalization rules for criteria 3 to 12 items in Table 2 used to calculate final dataset scores.

Item #	Description and Normalization Rule
3	Number of columns with dates: normalized from 1 to 5 using $NumCol = \frac{Val_o}{VMax_o}$, where Val_o is the original value to be normalized and $VMax_o = 5$.
4	Data aligned with the 17 UN Sustainable Development Goals: 1 if yes, 0 otherwise.
5	Data aligned with priority public policies (Multi-Year Plan): 1 if yes, 0 otherwise.
6–8	Difficulty due to number of records, columns, and process identification: normalized as $Diff = 1 - \frac{Val_o - VMin_o}{VMax_o - VMin_o}$.
9	Data with no noticeable problems: 1 if clean, 0 if problems are found.
10	No transformation needed from column to row: 1 if not needed, 0 otherwise.
11	File is an event log: 1 if yes, 0 otherwise.
12	Standardized data structure: 1 if yes, 0 otherwise.