Ensuring Data Quality in National Educational Databases: Insights from Brazil's Centralized Database of High School Students' Data

Abílio Nogueira Barros[a,b], Emanuel Marques Queiroga[c,b], Markson Rebelo Marcolino[d,b], Débora Barbosa Leite Silva[b], Diego Dermeval[b], André Lima[b], Leonardo Brandão Marques[b], Cristian Cechinel[d,b] and Thales Vieira[b*]

[a]Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE), Recife 52171, Brazil, abilionbarros@gmail.com

[b]Center of Excellence for Social Technologies (NEES), Universidade Federal de Alagoas, Maceió 57072, Brazil, {debora.silva, diego.matos, andre.lima, leonardo.marques, thales.vieira}@nees.ufal.br

[c]Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSul), Pelotas 96.015-360, Brazil, emanuel.marques@nees.ufal.br

[d]Centro de Ciências, Tecnologias e Saúde, Universidade Federal de Santa Catarina, Araranguá 88040, Brazil, cristian.cechinel@ufsc.br

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. This study investigates challenges in ensuring data quality within Brazil's national educational database, the Sistema Gestão Presente (SGP), and proposes solutions. Reliable and integrated data systems are critical for evidence-based policymaking, particularly in education. The SGP, designed to centralize student attendance and enrollment data, faces issues such as inconsistent data entry, logical errors, and systemic reporting anomalies. To address these challenges, a data-driven methodology inspired by the DMAIC framework was implemented, focusing on defining problems, measuring and analyzing data inconsistencies, improving processes through tailored solutions, and monitoring outcomes for continuous quality assurance. Seven case studies illustrate the results of this approach. These include resolving inconsistent enrollment dates, limiting multiple active enrollments per student, and ensuring consistency between disenrollment justifications and active statuses. Further, systemic anomalies, such as inflated attendance rates and implausibly high class hours reported at state levels, were identified and corrected through validation rules, training initiatives, and auditing mechanisms. These interventions reduced data inconsistencies, enhanced reliability, and improved system usability. The findings demonstrate how integrating validation mechanisms, improving data entry workflows, and fostering stakeholder collaboration can address large-scale data challenges. By ensuring the quality and integrity of educational data, the SGP enables more accurate insights into attendance patterns, dropout rates, and program effectiveness. These improvements lay the foundation for robust evidence-based policies and equitable educational outcomes across Brazil, highlighting the transformative potential of data quality assurance for public sector decision-making.

Keywords. Data quality, interoperability, data integrity, public policies, education.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.943

1. Introduction

The quality and integrity of educational data are foundational to developing evidence-based policies that drive meaningful improvements in education. This study explores the need for reliable educational data, focusing

on attendance records, and investigates common anomalies that can arise in large-scale datasets. Among the many data points collected in education systems, attendance data is a critical indicator of student engagement, resource allocation, and overall system performance. Accurate attendance records are more than just statistics. They tell how students interact with their learning environments and provide essential insights for shaping their futures. Without reliable data, policymakers risk basing decisions on flawed insights, leading to inefficiencies and missed opportunities for improving educational outcomes. Data reliability is paramount for fostering accountability and enabling informed decision-making (Jenkins & Duri, 2020).

This study draws on data from the *Sistema Gestão Presente* (SGP), a comprehensive database that tracks student attendance across Brazil, designed to integrate data from thousands of sub-national educational departments. The SGP is a critical tool for monitoring student engagement at a national level, providing detailed records of student attendance in classrooms across diverse educational settings. However, the scale and complexity of the SGP dataset bring unique challenges, such as data inconsistencies, varying collection methodologies, and discrepancies in reporting.

Data collection for large-scale educational systems often involves multiple stakeholders, including schools, districts, and state-level administrators. This decentralized approach creates unique challenges for interoperability, as data from various sources must be harmonized within a unified system. Differences in data formats, collection methods, and validation protocols can result in inconsistencies and errors that undermine the credibility of the entire dataset (Rocha et al., 2024).

Maintaining data quality goes beyond initial collection; it requires continuous verification throughout the data lifecycle. While validations at the point of entry are essential, they are insufficient to address the complexities of large and evolving datasets. A practical quality assurance framework must have ongoing monitoring, anomaly detection, and feedback mechanisms to identify and correct errors as they arise. These processes are crucial not only for maintaining the integrity of the data but also for building trust among the stakeholders who depend on it for informed decision-making (Redyuk et al., 2021; Taleb et al., 2021).

Reliable educational data, especially attendance records, is the backbone of effective policymaking. However, collecting accurate data on a nationwide scale is no simple task(Queiroga et al., 2024). Schools, districts, and state systems often use varied methods, making consolidating information into a single, coherent dataset challenging. By addressing these challenges and building systems that ensure data consistency, trustworthiness, and ease of integration, we can empower educators and policymakers with the insights needed to create impactful, student-focused policies and initiatives.

The present paper describes how data inconsistencies inside SGP were dealt with by adopting a data-driven approach to redesign the system based on post-implementation analysis and enhancements. The remainder of the paper is structured as follows. Section 2 describes the work focusing on data interoperability problems, while section 3 presents the educational context of SGP. Section 4 illustrates the methodology followed during the development of SGP to ensure data quality and integrity while describing the most important actors and stakeholders involved. Section 5 describes seven study cases to illustrate the data inconsistencies in the SGP and how they were resolved by adopting the presented methodology. Section 6 discusses the findings from the work, and section 7 concludes the paper.

2. Related Work

Data interoperability is a fundamental need in the public sector, encompassing different areas of governmental activity. As highlighted by Restrepo-Carmona et al., 2024, interoperability and the systematization of public databases can enhance transparency, accountability, and efficiency in the management of public funds. These examples demonstrate how interoperability transcends sectors, acting as a crucial pillar for the modernization and effectiveness of public administration.

Another relevant work on the importance of interoperability in the governmental context is Tavares and Bitencourt, 2022, which explores the relationship between interoperability and public policy evaluation within the scope of digital governance. Despite its significance, the implementation of interoperability faces significant barriers, such as the need for investments in technological infrastructure, personnel training, and cultural shifts within public institutions. These last two points, in particular, are directly tied to the success of data

usage, as public agents play a central role throughout the data lifecycle—from initial entry to final interpretation for decision-making. However, as noted by various studies in this field, the challenges faced are global and structural. The survey by Restrepo-Carmona et al., 2024, when analyzing the feasibility of interoperability in Latin American governments, identified recurring difficulties such as limited access to public data, lack of standardization, and the need for greater interinstitutional collaboration to integrate data efficiently. Additionally, infrastructure, governance, and organizational culture issues exacerbate these challenges.

Moving into the health sector, interoperability is also recognized as a critical factor for the reliability and quality of data. As noted by D'Amore et al., 2018, processes and tools aimed at cataloging and enhancing clinical data quality underscore the importance of well-defined standards. Similarly, Valdés and Lecaros, 2024 proposes necessary actions for creating biobanks in the Latin American context, aiming to strengthen research and produce high-quality scientific evidence on a global scale.

Interoperability plays a crucial role in Brazil's healthcare system, particularly for data quality and system integration. According to the systematic review by (Barbalho et al., 2022), 78.58% of analyzed studies highlight interoperability between systems as essential for improving patient care. Their research identified 31 national health information systems operating in primary care between 2013-2018, with significant fragmentation challenges. This fragmentation became more evident in the implementation of e-SUS Primary Care Strategy, as shown by (Coelho Neto et al., 2021), where only 12 of 31 systems achieved complete integration with e-SUS interfaces. In comparison, 15 systems showed no integration at all. The study revealed a stark disparity between systems managed by different government departments - notably between the Health Care Secretariat (SAS) and Health Surveillance Secretariat (SVS), where SVS systems showed minimal integration despite having similar numbers of systems in use. Similarly, (Vasques et al., 2024) propose a maturity framework for evaluating interoperability in Brazilian healthcare institutions, noting that despite SUS protocols established in 2011 to coordinate information from over 6,000 healthcare providers, private institutions often develop independent systems that minimize contribution to the broader Brazilian health network. This fragmentation persists despite initiatives like the National Health Data Network (RNDS), which impact data quality and limit the potential for integrated healthcare delivery.

Entering the educational domain, several studies explore interoperability and data adequacy for implementing and improving educational systems. For instance, Macedo de Andrade et al., 2022 evaluates a program's impact on encouraging educational data use by school actors. This study employed a quasi-experimental Differences-in-Differences (DiD) model to analyze the effects of the *Escolas em Foco* program. The methodology combines advanced statistical techniques to measure the program's impact while addressing the importance of data interoperability to enhance student learning. The analysis was conducted using various datasets, highlighting how the effective use of such information can support improvements in school performance.

The work of BOZ and SIMSEK, 2022, the interoperability of information systems within Turkey's Ministry of National Education (MoNE) is explored in detail. The article emphasizes that MoNE is the country's largest public institution, serving 18 million students and employing over one million teachers. To manage this robust structure, the ministry operates several information systems. During the study, metrics were identified to evaluate interoperability and strategies to improve intra-system communication. The work stresses that a stronger correlation between data is essential for more in-depth analyses, promoting advancements in the use of educational information.

In the Brazilian context, Queiroga et al., 2024 addresses the integration of the SAEB (National System of Basic Education Evaluation) and School Census databases. This study connects information about student performance and school conditions, enabling comprehensive analyses and identifying factors that impact educational equity in Brazil. The integration of these databases allowed researchers to go beyond isolated analyses, generating deeper *insights* into the relationships between school infrastructure and educational outcomes. The study reinforces the importance of reliable and interconnected data to provide a current overview and support public policies to improve Brazilian education.

Once the importance of data interoperability has been analyzed, data quality is another crucial factor for using these data. The study by Wijayanti et al., 2018 evaluates data quality in higher education, focusing on the Statistics Institute of Indonesia. This study used a quantitative approach to measure data quality, followed by a qualitative analysis of the results. Direct queries in SQL (Structured Query Language) were conducted in the integrated information system database. The evaluation revealed that the data quality at the Statistics Insti-

tute does not meet governmental requirements. Observing the completeness pillar, only 21.15% of business rules were met, and regarding validity, 57.90% of business rules were fulfilled, with issues found related to inconsistent data formats and types.

To solve the same problem of ensuring data quality in integrating university data, the study at Universitas Amikom Purwokerto Wahid et al., 2024 implemented a data warehouse. This data warehouse meets reporting requirements and promotes a culture of high-quality data management, where data is accessible, accurate, and useful for decision-making and strategic planning. Additionally, Du, 2021 described a model that seeks to define a theoretical and methodological framework for classifying, collecting, analyzing, and evaluating educational data quality, aiming to improve the management and utilization of these data. The study addresses the importance of a data quality management system in universities, encompassing quality assurance, process management, and data quality evaluation. The proposed model also emphasizes the need for a data quality evaluation system to ensure that collected data is accurate and complete. Although more focused on academic systems, these examples highlight the importance of data quality in connected contexts, demonstrating that efforts must be systematic and continuous to ensure the accurate and correct use of these data.

3. Educational context - The Sistema Gestão Presente (SGP)

The *Sistema Gestão Presente* (SGP)[1] is a new tool developed by Brazil's Ministry of Education (MEC) to enable comprehensive monitoring of students in basic education. This system is designed to centralize student enrollment and attendance tracking information, providing a sustainable and integrated platform for educational networks to submit data. Its implementation aims to streamline data management processes while enhancing accessibility and reliability across the education sector.

To accommodate the diverse realities of educational networks, the SGP allows for two formats of information submission: through APIs for educational networks with integrated systems and via data spreadsheets for those lacking advanced technological infrastructure. Figure 1 illustrates the data flow and integration within the SGP, highlighting its role in managing student data for Brazil's educational networks.

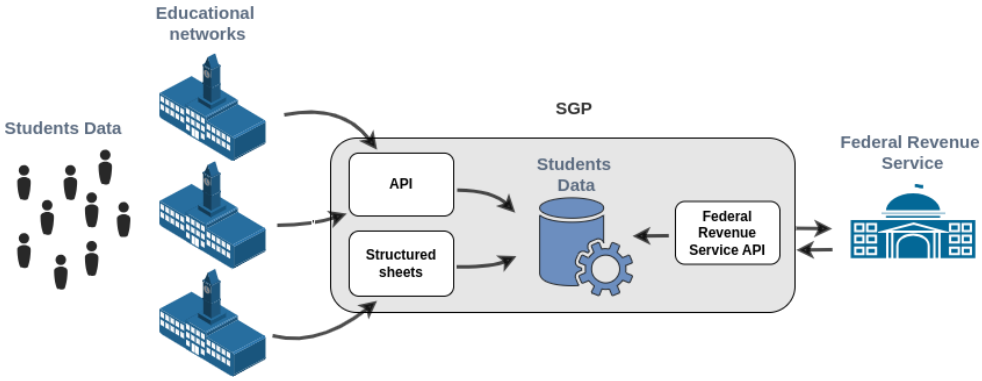

Fig. 1 – Sistema Gestão Presente.

The figure shows that the process begins with student data collected from schools across various educational networks. This data is transmitted to the SGP via structured spreadsheets or an API, depending on the technological capabilities of the institutions. Once in the SGP, the student data is stored in a centralized database, which can be processed and utilized. The system also integrates with the Federal Revenue Service through an API, enabling a two-way data exchange to support compliance, verification, and analysis needs.

This flexibility aims to enhance the system's reach and efficiency, ensuring that schools with robust systems and those with more limited resources can participate inclusively. With the consolidated database, studies can be conducted, and techniques can be applied to identify patterns and generate alerts about critical issues such as school dropout rates, student attendance patterns, or abrupt changes in attendance behavior. For instance, students with historically high attendance rates who suddenly show significant declines may indicate situations requiring closer monitoring by the school. This oversight allows for proactive interventions to address

[1] https://gestaopresente.mec.gov.br/

problems and promote student retention in the school environment (Macarini et al., 2020; Queiroga et al., 2022).

Implementing a system to support educational management in a country with continental dimensions, such as Brazil, is a significant challenge. In 2023, the National Institute for Educational Studies and Research Anísio Teixeira. (INEP) reported 47.3 million enrollments distributed across 178,500 educational entities in BrazilBrasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2024, covering the entire national territory. Given the magnitude of this challenge, obstacles related to integrating various education networks are expected, considering their specific educational characteristics, such as differences in the number of school days, workloads, and regional particularities. These challenges were reflected in the data entered during the system's initial months of use. Errors or inconsistencies were identified, highlighting initial difficulties in adaptation by educational institutions.

The following section will address the anomalies detected and reported during this period, aiming to highlight the actions taken to improve the quality and accuracy of the data entered into the system.

4. Methodology for data quality assurance

This section describes the methodology currently followed by the SGP team to ensure the quality and integrity of data inside the system. Subsection 4.1 describes the main actors involved in this process, and 4.2 depicts the different steps followed during the process.

4.1. Working teams and stakeholders

Leading a project of great significance, such as the development and maintenance of the SGP, requires the integrated effort of multidisciplinary teams performing different roles within the organizational structure. This collaboration is essential for the system to meet the complex demands of the Brazilian educational context. At the project's core is project management, tasked with planning, coordinating, and ensuring the completion of each stage within the established deadlines and objectives while adapting to demands arising from interactions with external stakeholders. Developers, in turn, are responsible for implementing technical features, ensuring system robustness, user interface efficiency, database integrity, and the continuous refinement of business rules. Simultaneously, the data quality team monitors and validates the consistency and integrity of the data, ensuring its reliability for internal and external applications, such as future studies and analyses.

Additionally, the team focuses on the public policy perspective and strategically aligns project activities with societal priorities. This team works to maximize the system's impact, manage external demands, prepare manuals, organize training sessions, and mediate potential conflicts. Collaboration with external stakeholders is equally indispensable. The Ministry of Education, as the central coordinator, sets strategic guidelines and ensures alignment with national public policies. Meanwhile, state and municipal education departments are critical in implementing the system locally. They act as a bridge between the SGP and educational institutions, providing data, offering initial support, and serving as schools' first point of contact. Figure 2 depicts the most important actors involved in the process of ensure data quality inside SGP.

This integration between internal and external teams ensures that the SGP operates efficiently and is aligned with its users' real needs. As a result, the system directly contributes to quality educational management that is equitable and responsive to regional and national challenges.

4.2. Workflow

The methodology for data quality assurance follows a data-driven approach to problem-solving and optimization with a similar structure and purpose of lean DMAIC (Define, Measure, Analyze, Improve, Control) Six Sigma framework (Srinivasan et al., 2016). The process begins with problem definition ("Define" phase in DMAIC), focusing on clarifying hypotheses and research questions to set the foundation for analysis. In the design and prototyping phase ("Measure" and "Analyze" stages from DMAIC), data is systematically examined using statistical methods, machine learning, and visualization to identify and understand anomalies. The implementation phase ("Improve" phase from DMAIC) translates analytical findings into actionable solutions such as new business rules, training, and auditing to resolve identified issues. Lastly, the evaluation

Fig. 2 – Actors involved on SGP's data quality assurance.

and learning phase mirrors the "Control" phase, emphasizing monitoring implemented solutions, ensuring their effectiveness, and capturing insights for continuous improvement. Figure 3 gives an overall view of the methodology cycle and its actors.

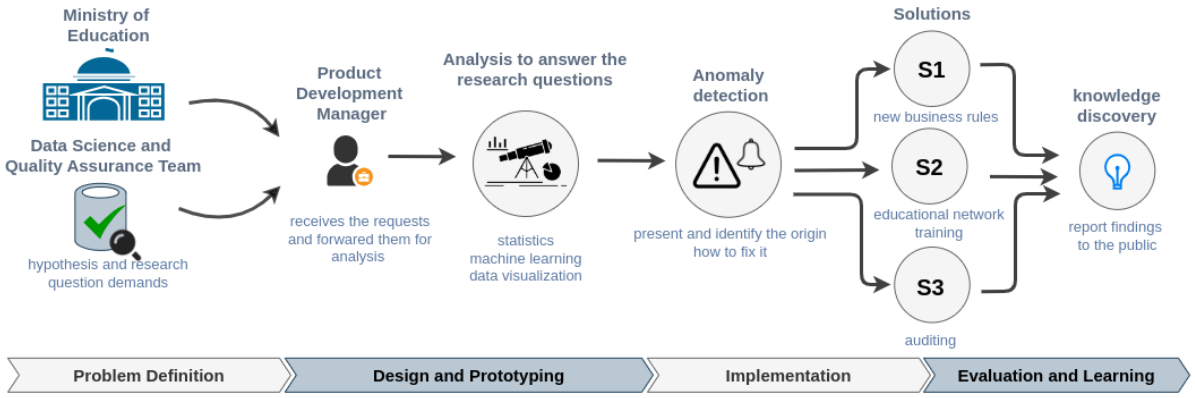

Fig. 3 – Overview of the data quality assurance process.

The figure shows that the process starts with the Ministry of Education or other stakeholders submitting hypotheses and research questions to the Data Science and Quality Assurance Team. The **problem definition** phase identifies and categorizes any inconsistencies in the data. This includes raising concerns and forming hypotheses to address incorrect data inputs or calculation errors. It's also a collaborative process involving stakeholders in understanding the scope and root causes of the problems entirely. The Product Development Manager plays a key role here, providing a bridge to ensure these issues are forwarded for deeper analysis. At this stage, the data is reviewed in detail to uncover patterns, outliers, and areas that need further investigation. This analysis lays the groundwork for the next steps and relies on technical expertise and a strong understanding of the educational context. From this work, two main types of hypotheses typically emerge:

- Team-Driven Hypotheses: These are based on patterns, anomalies, or inconsistencies uncovered through internal analysis.
- Externally Requested Hypotheses: These come from higher-level teams or the Ministry of Education (MEC) and address specific institutional priorities.

Once the hypotheses are formed, they are critically analyzed to ensure relevance and impact. This involves combining technical knowledge with an understanding of educational data and business rules, ensuring the investigations are well-targeted and aligned with the broader goals.

Data is then analyzed in the **design and prototyping** phase to address the research questions. Here, some techniques are used, such as statistical methods (quantitatively measuring and evaluating data), machine

learning (to detect more complex patterns or anomalies that are challenging to identify manually), and data visualization techniques (that assist in clearly and comprehensively communicating findings). Next, anomalies are identified, their causes traced, and potential solutions proposed. Tailored solutions are developed for each specific case. For example, new business rules are prototyped and tested on subsets of data, while training programs and audit strategies are designed iteratively to incorporate feedback from stakeholders.

The **implementation phase** focuses on implementing and deploying these solutions incrementally to minimize disruptions. Business rules were implemented in the system with automated validation, training sessions were delivered to users, and auditing processes were institutionalized to ensure ongoing data integrity. Three main types of outcomes have already been implemented according to the nature of each issue. These solutions included:

1. (S1) creation of new business rules within SGP - addressing scenarios where inconsistencies required new rules to be implemented at the system level to prevent similar issues in the future;
2. (S2) Training within the educational network - Cases involving manual input or network-related errors were mitigated through user training programs aimed at reducing future occurrences and
3. (S3) Data Auditing - Recommendations to the managers and government stakeholders to audit inconsistencies that were not resolved to ensure data reliability and integrity

Finally, the process concludes with **evaluation and learning**, where the outcomes of the implemented solutions are evaluated. Metrics such as the inconsistencies of anomalies, user feedback, and system performance are analyzed to measure the effectiveness of the interventions. Lessons learned are documented to refine future processes and ensure continuous improvement and shared publicly to foster knowledge discovery.

5. Case studies

To illustrate the methodology presented in the previous section, we present seven cases developed for eight months related to the data quality assurance process inside SGP. Each case has particularities that help to understand the nuances and complexities of data quality assurance inside a national educational system that receives information from all nationwide educational networks. As this process is cycled and is still under development, some solutions for specific cases are presented as suggestions for implementation and still need further evaluation.

5.1. Case 1: Inconsistent Enrollment Dates

Problem Definition: Inconsistent start dates were observed in the student's enrollment data. These inconsistencies suggest incorrect data entry, which can skew analysis when dates are used as filters or reference points for decision-making. To identify the scope of the problem, we executed a query that excluded the *starting date* condition, capturing all enrollment records regardless of their start dates. The analysis revealed a range of anomalous start years, from as early as 1900 (implying a 125-year-old high school student) to as late as 2029 (suggesting enrollment 4 years into the future). The most frequent inconsistencies were found in the years immediately preceding 2024. For example, 382,221 enrollments were recorded in 2022 and 153,407 in 2023. Outlier years, such as 1900, 1985, and 2029, also appeared in the dataset but with minimal occurrences (one enrollment each). Other years, such as 2016 (222 enrollments), 2018 (563 enrollments), and 2021 (3,659 enrollments), suggested a gradual increase in data reporting errors over time. These inconsistencies likely resulted from manual data entry errors or the lack of validation rules to ensure accurate recording of start dates. It is important to highlight that SGP is dealing with information from students enrolled in high school education, so dates such as 2016 or 1018 are considered not coherent with the expected date ranges.

Design and Prototyping: To address this issue, we proposed implementing validation rules to restrict start dates to a predefined range, with the current year as the last valid year for enrollment records. A prototype validation mechanism was designed to flag and reject entries with start dates outside the acceptable range during data input. This prototype was tested on a subset of historical data to evaluate its effectiveness and to identify potential edge cases, such as valid future or historical enrollments. In addition to technical solutions, systemic measures were proposed to minimize recurring errors. Training sessions are suggested to be developed for data entry staff to emphasize the importance of accurate data recording and to reduce manual errors.

Furthermore, a reporting dashboard was designed to highlight anomalous dates, enabling administrators to review and correct outliers in real-time.

Implementation: The validation mechanism was implemented in the application layer, and the date entry interface was updated with a date picker that restricted inputs to valid years, such as 2024. The team suggested guidelines explaining the validation rules for stakeholders to ensure a smooth transition. Additionally, historical data containing invalid start dates was retrospectively cleaned, with corrections prioritized for the most critical discrepancies. This combination of preventive measures and corrective actions ensured a robust implementation.

Evaluation and Learning: The solution's success is evaluated using several metrics. These include monitoring the reduction of invalid start dates in new records and assessing the efficiency of the reporting dashboard in flagging and correcting errors. User feedback is also being collected to evaluate the usability of the updated date entry interface and the clarity of the validation rules. The findings from this case underscore the importance of integrating validation mechanisms into data entry processes to improve overall data quality. The lessons learned are applied to other system areas with similar data validation challenges. These iterative improvements demonstrate a commitment to maintaining accurate and reliable data for analysis and decision-making.

5.2. Case 2: Multiple Active Enrollments for a Student

Problem Definition: After defining the criteria for an active enrollment, we identified cases where individuals had multiple active enrollments associated with their ID. While some instances of dual active enrollments are expected and valid, such as students simultaneously attending regular and technical schools, other cases raised concerns. Specifically, we focused on individuals with more than two active enrollments, which deviates from the expected norm. Our analysis revealed 138,747 individuals with more than two active enrollments at the time of the query. The presence of such anomalies indicates potential issues within the system, including errors in the registration process that allow duplicate or conflicting records, the absence of validation mechanisms to limit the number of active enrollments per individual, or possible misuse of the system—whether intentional or unintentional—resulting in inflated enrollment counts.

Design and Prototyping: To address this issue, we proposed implementing system-level validations to restrict the number of active enrollments associated with a single student ID to a maximum of two, accounting for valid scenarios such as simultaneous enrollment in regular and technical schools. A prototype validation rule was designed and tested on historical data to identify and flag records exceeding the defined threshold. In addition to this technical solution, we developed guidelines to clarify the conditions under which dual enrollments are permissible, ensuring alignment with institutional policies. These guidelines were shared with stakeholders to understand legitimate enrollment scenarios better and standardize data entry practices.

Implementation: The validation mechanism was integrated into the enrollment system to prevent the creation of new records that exceed the maximum allowable number of active enrollments—this involved frontend restrictions during the registration process and backend validations at the database level. A cleanup process was initiated for existing records to address individuals flagged with more than two active enrollments. This included engaging with educational institutions to verify and resolve duplicate or conflicting records. Training sessions were conducted for registration staff to support these efforts, emphasizing the importance of adhering to the validation rules and accurately categorizing dual enrollments. A reporting mechanism was also introduced to monitor new instances of multiple active enrollments, allowing for continuous oversight and prompt resolution of anomalies.

Evaluation and Learning: The success of the implemented solution is being evaluated through a reduction in the number of individuals with more than two active enrollments in subsequent data analyses. Additionally, the reporting mechanism tracks compliance with the validation rules and detects any remaining anomalies in real time. Feedback from registration staff and institutional stakeholders is being collected to assess the guidelines' clarity and the usability of the system improvements. This case highlights the importance of establishing precise validation mechanisms and robust monitoring processes to maintain data integrity. The insights gained from addressing this issue are applied to other system areas to prevent similar anomalies and ensure that the data accurately reflects real-world conditions.

5.3. Case 3: Enrollments with Justification for Disenrollment and Active Status

Problem Definition: In the enrollments table, the column `co_tipo_justificativa_desmatricula` records the reason for disenrollment. This field only contains a value if the enrollment is deactivated. However, our analysis identified cases where enrollments were simultaneously marked as active (`st_matricula_ativa = true`) and had a justification for disenrollment recorded. At the time of analysis, 26,106 enrollments exhibited this inconsistency. This issue suggests potential data entry or synchronization errors within the system. The inconsistencies likely result from incorrect updates to the enrollment status, where records remain active despite a justification for disenrollment being provided. Other possible causes include miscommunication between system modules responsible for managing enrollment statuses and disenrollment reasons and the lack of validation checks to prevent conflicting information in these fields during data entry. This inconsistency affects data reliability, particularly in metrics that accurately report active versus inactive enrollments and analyze disenrollment trends, which rely on precisely categorizing enrollment statuses.

Design and Prototyping: To address this issue, we proposed implementing validation rules to ensure that a justification for disenrollment can only be recorded if the enrollment status is marked as inactive. These rules were designed to operate at both the application and database levels to prevent conflicting information from being entered. A prototype validation mechanism was developed and tested on historical data to identify the extent of the issue and ensure that the rules functioned correctly. In addition to the validation rules, we proposed reviewing and redesigning the data entry workflows to reduce the likelihood of errors during status updates. This included automating clearing disenrollment justifications when enrollment is reactivated and enhancing system logs to capture changes to the enrollment status and associated justification fields.

Implementation: The validation mechanism was implemented to enforce consistency between the enrollment status and the justification for disenrollment fields. In the application layer, updates to the status field were synchronized with the justification field, ensuring that conflicting values could not be saved. At the database level, constraints were added to restrict the recording of a justification for disenrollment to inactive enrollments only. A system-wide data cleanup was performed to resolve existing inconsistencies. This involved reviewing the 26,106 flagged records and correcting them either by deactivating the enrollment or clearing the justification field, depending on the circumstances. Registration staff is also advised to be provided with guidelines detailing the correct process for handling enrollment status updates to ensure compliance with the new rules.

Evaluation and Learning: The solution's effectiveness is evaluated by continuously monitoring new enrollment records to ensure no further inconsistencies occur. Metrics include the number of new flagged records, the accuracy of active versus inactive enrollment reports, and the consistency of disenrollment trend analyses. Feedback from system users and administrators is suggested to be collected to assess the updated workflows' usability and the clarity of the validation rules. This case highlights the importance of validation mechanisms and system automation in maintaining data integrity.

5.4. Case 4: Analysis on the Application of 100% Attendances at the State Level

Problem Definition: Building upon the resolution of fundamental data inconsistencies, we expanded our analysis to investigate more complex anomalies, focusing specifically on the accuracy of attendance reporting at the state level. The aim was to assess whether all states correctly reported student attendance data and identify potential patterns or inconsistencies in applying 100% attendance rates. To achieve this, we calculated the percentage of full attendances relative to the total informed attendances for each State and month. This analysis enabled us to identify the State with the highest rate of full attendance across the analyzed period. The results revealed that State A reported the highest rate of full attendance in July 2024, with 104,937 enrollments (out of 140,859 informed enrollments) marked as having 100% attendance. This corresponded to a rate of 74.5%, the highest among all states during this timeframe. Such high rates of attendance of 100% could indicate genuine trends of high attendance, potentially reflecting efficient educational policies or practices. However, they could also result from overreporting or inaccuracies in data submission by institutions, raising questions about the reliability of attendance reporting.

Design and Prototyping: To investigate the validity of these findings, we designed an analysis framework to detect potential overreporting patterns and validate attendance data at the state level. This involved cross-referencing attendance data with other indicators, such as dropout rates and institutional capacity, to evaluate

whether the reported rates of 100% attendance were consistent with different metrics. Prototypes of visual dashboards were developed to present the attendance data in an accessible format, highlighting states and months with unusually high rates of full attendance. These dashboards allowed stakeholders to identify and investigate potential anomalies quickly.

Evaluation and Learning: The evaluation phase focused on monitoring flagged cases of potential overreporting and assessing the impact of the new reporting guidelines on attendance data quality. Metrics such as the number of flagged anomalies, feedback from state administrators, and improvements in attendance consistency were used to evaluate the effectiveness of the implemented measures. This case underscores the importance of cross-referencing attendance data with other indicators to ensure reliability. The findings have provided valuable insights into attendance reporting practices at the state level and have informed the development of additional validation mechanisms to maintain data integrity in future analyses.

5.5. Case 5: Hours Studied Greater than Hours Offered

Problem Definition: During an exploratory data analysis of the database, an inconsistency was identified where the reported hours studied exceeded the number of hours offered. This issue undermines the data's reliability, which could affect decision-making processes and policy outcomes. Queries were executed to identify the scope of the problem, and inconsistencies were flagged across three months. It is important to highlight that both information (hours studied and hours offered) are reported from the same source (educational network).

- In February 2024, there were 7,119 enrollments in State A, primarily in State schools, where the hours studied were greater than those offered.
- In March 2024, a similar issue was detected in State B, with 15 enrollments, all in State schools.
- By April 2024, 106 enrollments in State C were identified with the same inconsistency, predominantly in Federal schools.

The root causes of this issue likely include manual data entry errors leading to overreported hours studied, miscommunication between system modules responsible for calculating hours, and potential discrepancies in data aggregation processes at the institutional or regional level. The inconsistencies were reported to the product owner to address this issue, who prioritized a solution to mitigate its impact and ensure data consistency.

Design and Prototyping: The proposed solution involved implementing a business rule to prevent hours studied from exceeding hours offered. This rule was designed to operate at the database level, integrating validation mechanisms to stop invalid entries during data input and flagging existing anomalies for manual review. In addition to the technical solution, measures were proposed to address systemic issues, such as improved training for staff responsible for data entry and introducing auditing processes to monitor data integrity, particularly in regions with a high incidence of anomalies.

Implementation: The business rule deployment was planned carefully to minimize disruption. The rule was integrated into the application layer (via form validation) and the database (using triggers or constraints). Users and stakeholders were informed about the change, and guidelines were provided to ensure smooth adoption. A correction strategy was devised to address existing anomalies, prioritizing regions and school types most affected by the issue.

Evaluation and Learning: The success of the implementation is being measured through the reduction of inconsistencies over subsequent months and the Frequency and distribution of rule triggers. It is also suggested that feedback be collected from users and administrators regarding the rule's usability and effectiveness.

5.6. Case 6: Enrollments by State with More than 300 Studied Class Hours

Problem Definition: In our analysis, we identified cases where students were reported to study more than 300 class hours within a month. This threshold was selected as an inconsistency, representing an average of over 10 hours of study per day — an unlikely scenario given standard academic schedules and considering a maximum of 22 days per month (220 class hours monthly plus 80 extra hours). Such cases raise questions

about the accuracy of the reported data and the underlying processes involved in calculating hours studied. This threshold can be considered quite conservative, and similar analyses could also consider lesser and more conservative thresholds.

Our investigation revealed significant findings across several states. State D reported the highest number of such enrollments in February 2024, with 2,891 enrollments exceeding the 300-hour threshold. Similarly, State E showed consistently high numbers, with 2,221 enrollments in February 2024 and 2,087 enrollments in March 2024. Other notable occurrences included State F, which reported 2,466 enrollments in November 2024, and State G, with 1,503 enrollments in February 2024. These patterns highlight potential systemic issues in the reporting of studied class hours. These inconsistencies suggest possible causes such as data entry errors, where hours studied were inaccurately recorded, miscommunication or discrepancies in how hours are calculated or aggregated at the institutional level, and the absence of validation mechanisms to flag and correct unfeasible data entries during the reporting process.

Design and Prototyping: To address these anomalies, we suggested implementing validation rules to flag cases where reported class hours exceed a predefined threshold. The prototype for these rules included automated checks to identify and mark entries exceeding the 300-hour limit for review. Historical data was used to simulate the behavior of these checks, allowing us to evaluate their effectiveness in identifying potential errors without hindering legitimate data. In addition to validation mechanisms, we designed a feedback loop involving institutional administrators. This process allowed flagged data to be reviewed collaboratively, ensuring that the identified anomalies were corrected or justified. Training materials and guidelines were developed to clarify how class hours should be calculated and reported, reducing the likelihood of recurring errors.

Implementation: The validation rules were integrated into the reporting system, ensuring entries exceeding the 300-hour threshold were flagged automatically for manual review. This integration occurred at both the data entry and aggregation stages, providing multiple validation layers to prevent unfeasible data from being finalized in reports. A retrospective cleanup was also conducted to address the anomalies identified in historical data. States with the highest occurrences of anomalies, such as State D and State E, were prioritized in this process. Communication channels were established with educational institutions in these states to resolve discrepancies and implement corrective measures.

Evaluation and Learning: The evaluation phase focuses on assessing the impact of the implemented measures. Metrics include a reduction in the number of enrollments reported with more than 300 class hours and improvements in the overall accuracy of reported class hours. Feedback from institutional administrators is suggested to be collected to refine the validation rules and ensure their practicality in day-to-day reporting. This case underscores the importance of robust validation mechanisms and collaborative review processes in maintaining data accuracy. By addressing the root causes of these anomalies, the system is better equipped to handle similar issues in the future, ensuring the reliability of data used for analysis and decision-making. Lessons learned from this case are being applied to other areas of data reporting to promote consistency and prevent the propagation of errors.

5.7. Case 7: Monthly Class Offered Hours Higher than 300

Problem Definition: An additional anomaly was identified in the data, involving cases where schools reported offering more than 300 class hours per month for enrollment. This threshold is considered unfeasible, as it implies an average of over 10 hours per day, which is highly unlikely under regular academic schedules. Such cases raise questions about the accuracy of the reported data and the processes involved in calculating the number of hours offered. Our analysis revealed several notable findings. State H reported the highest occurrence of such cases in February 2024, with 3,831 enrollments exceeding the 300-hour threshold. State I also exhibited consistently high numbers, including 2,221 enrollments in February 2024 and 2,087 enrollments in March 2024. Other states with significant occurrences included State J, which reported 1,503 enrollments in February 2024, and State K, with 2,898 enrollments in November 2024. These patterns suggest systemic issues that may lead to overreporting or inaccuracies in the reported hours offered. The anomalies could stem from data entry errors, such as inaccurately recording the number of hours offered, discrepancies in how hours are calculated or aggregated at the institutional level, or the absence of validation checks to flag and correct unfeasible data entries during reporting.

Design and Prototyping: To address these anomalies, we proposed developing validation rules to flag in-

stances where the reported hours offered exceed the 300-hour threshold. The validation process included automated checks to identify and mark records exceeding this limit for manual review. Prototyping used historical data to simulate these rules' behavior and ensure they effectively captured anomalies without hindering legitimate data entries. In addition to the validation rules, we suggested developing guidelines to standardize the reporting hours offered. These guidelines should emphasize the importance of accurate data entry and provide clear instructions on how hours should be calculated and aggregated. Training sessions were also suggested to educate institutional staff about these standards and the role of validation mechanisms in ensuring data integrity.

Implementation: The validation rules were integrated into the reporting system, with automated checks applied during data entry and aggregation. These checks flagged entries exceeding the 300-hour limit, ensuring they were reviewed, corrected, or justified. A retrospective cleanup process was also carried out to address historical anomalies, prioritizing states with the highest occurrences, such as State H and State I. Communication channels were suggested to be established with institutional administrators to facilitate the resolution of flagged inconsistencies and to ensure that corrective measures were effectively implemented. The reporting system was also enhanced to provide real-time feedback to users entering data, reducing the likelihood of future errors.

Evaluation and Learning: The effectiveness of the implemented solutions is being evaluated by monitoring the reduction in the number of enrollments exceeding the 300-hour threshold in subsequent reporting periods. Metrics such as the frequency of flagged anomalies, the resolution time for flagged cases, and feedback from institutional administrators are being used to assess the impact of these measures. This case highlights the importance of establishing robust validation mechanisms and fostering collaboration with institutional stakeholders to ensure accurate data reporting. The insights gained from addressing these anomalies are being used to refine the validation processes further and to apply similar approaches to other areas of data reporting. Through iterative learning, the system continues to evolve, ensuring the reliability and accuracy of reported data for analysis and decision-making.

6. Discussion

The seven cases of data inconsistencies previously presented can be broadly categorized into three main groups according to the types of errors found: Input Data Errors, Logical Inconsistencies, and Systemic or Aggregated Inconsistencies. Input Data Errors arise from manual data entry mistakes or the absence of proper validation mechanisms, as seen in cases like inconsistent enrollment dates and study hours exceeding offered hours. These errors rely primarily on validation and automated solutions to prevent mistakes at entry, together with training and educating users. Logical Inconsistencies occur when conflicting or unrealistic data is reported, such as students having multiple active enrollments or simultaneous active statuses with disenrollment justifications. Logical inconsistencies are mainly resolved through validation to enforce consistent logic and monitoring to detect real-time conflicts. Lastly, Systemic or Aggregated Anomalies reflect broader reporting or systemic issues, including implausible state-wide 100% attendance rates or unrealistic monthly study and offered hours that exceed feasible academic schedules. These errors require monitoring to flag irregular patterns and training to ensure adherence to reporting standards.

To foster data-driven public policies in education in Brazil, data infrastructure must be improved on the three fronts described by (Sellar, 2015), which are: *scales* across which the infrastructure is being integrated, the *scope* of the data being created, and the *explanatory power* enabled by the data infrastructure. SGP can be considered the first nationwide initiative focused on developing a unified database in the educational context that reaches the granularity of the data at a personal level (student level). In Brazil, public educational institutions are managed by different state entities according to their educational level. For instance, the Federal Government typically manages public federal universities, while State Governments manage high school education and, in some cases, Upper middle schools, and Municipalities manage Lower and Upper Middle schools and Early Childhood Education. Considering this administrative division, each public organization is also responsible for maintaining, securing, and managing its data (Filgueiras & Lui, 2023). This makes it very difficult to have an integrated student attendance centralized database, which is key for advancing evidence-based and data-driven policy capabilities in the country(Queiroga et al., 2024; Williamson, 2019).

Once that interoperability issues are overcome and data quality integrity is improved inside SGP, it becomes feasible to foster knowledge discovery inside the system to promote assertive public policies. For instance,

soon, it will become possible to evaluate student retention rates and compare the outcomes of beneficiaries and non-beneficiaries of social programs, enabling targeted interventions to address challenges in academic progression. Data integration also allows for analyzing school attendance patterns before and after specific events or seasons, such as natural disasters or agricultural cycles, helping policymakers design strategies to mitigate disruptions. Moreover, precise definitions for identifying dropouts, supported by accurate and comprehensive data, ensure precise tracking of students' educational trajectories and facilitate the development of effective measures to reduce dropout rates. Advanced techniques, such as explainable AI (xAI) and clustering, can help to provide a deeper understanding of attendance behaviors, highlighting high and low-risk groups for intervention. Integrated and high-quality data also support the development of tools to monitor the impact of educational programs, enabling continuous evaluation, refinement, and improved policy outcomes. At the regional level, data integration facilitates global and state-level clustering, allowing for the creation of tailored policies that address regional disparities and promote equity. Specific public educational program analyses become actionable through comprehensive datasets, offering insights into program effectiveness and identifying areas for improvement. Furthermore, SGP data may enable comparisons of economic indicators, such as resource allocation relative to population demographics, optimizing public investments in education, and enhancing outcomes across various contexts.

7. Final Remarks

The results from this study demonstrate the transformative impact of addressing data inconsistencies within Brazil's Sistema Gestão Presente (SGP). The application of a structured, data-driven methodology effectively helped to resolve critical challenges such as inconsistent enrollment dates, redundant active enrollments, and systemic reporting anomalies. Notably, introducing tailored validation mechanisms, automated error detection, and the proposal of enhanced training initiatives can significantly improve data integrity and usability. By systematically identifying and correcting anomalies, the SGP interactively enhances its reliability as a national educational database and establishes itself as a model for continuous quality assurance in large-scale public systems.

These findings provide valuable insights for other organizations aiming to improve the data quality of their systems. By adopting a similar approach that combines robust validation protocols, stakeholder engagement, and ongoing monitoring, other systems can achieve greater consistency and accuracy in their datasets. Additionally, the lessons from SGP's case studies highlight the importance of integrating technical solutions with user training and feedback loops. These practices help to ensure data quality and foster stakeholder confidence, enabling better-informed decisions and more effective public policies.

Acknowledgments

This work was supported by the Ministry of Education (MEC), Brazil, under Grant TED13914. The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript. Additionally, the funders had no influence on the interpretation of data and the final conclusions drawn.

References

Barbalho, I. M., Fernandes, F., Barros, D. M., Paiva, J. C., Henriques, J., Morais, A. H., Coutinho, K. D., Coelho Neto, G. C., Chioro, A., & Valentim, R. A. (2022). Electronic health records in brazil: Prospects and technological challenges. *Frontiers in Public Health*, *10*, 963841.

BOZ, M., & SIMSEK, I. (2022). Analysis of education management information systems of the ministry of national education in terms of interoperability. *Journal of Qualitative Research in Education*, (32). DOI: https://doi.org/10.14689/enad.32.1702.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2024). *Censo da educação básica 2023: Notas estatísticas*. Inep.

Coelho Neto, G. C., Andreazza, R., & Chioro, A. (2021). Integration among national health information systems in brazil: The case of e-sus primary care. *Revista de saude publica*, *55*, 93.

D'Amore, J., et al. (2018). Interoperability progress and remaining data quality barriers of certified health information technologies. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2018*, 358–367.

Du, J. (2021). Research on the construction of educational data quality model based on multiple constraints model. *2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 363–367.

Filgueiras, F., & Lui, L. (2023). Designing data governance in brazil: An institutional analysis. *Policy Design and Practice*, *6*(1), 41–56.

Jenkins, M., & Duri, J. (2020). *The role of evidence, data and research findings in promoting integrity in education* (tech. rep.) (This document is part of the Transparency International Anti-Corruption Helpdesk series.). Transparency International. Berlin, Germany. https://knowledgehub.transparency.org/assets/uploads/kproducts/The-role-of-evidence-data-and-research-findings-in-promoting-integrity-in-education.pdf

Macarini, L. A. M., dos Santos, H. L., Cechinel, C., Ochoa, X. O., Rodés, V., Casas, A. P., Lucas, P. P., Maya, R., Alonso, G. E. A., & Díaz, P. (2020). Towards the implementation of a countrywide k-12 learning analytics initiative in uruguay. *Interactive Learning Environments*, *28*(2), 166–190. DOI: https://doi.org/10.1080/10494820.2019.1636082.

Macedo de Andrade, F., Ceneviva, R., & Koslisnki, M. C. (2022). Escolas em foco: A avaliação de impacto do programa de'políticas públicas baseadas em evidências' da rede municipal de ensino do rio de janeiro. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas/Arquivos Analíticos de Políticas Educativas*, *30*.

Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. *Information*, *13*(9). DOI: https://doi.org/10.3390/info13090401.

Queiroga, E. M., Siqueira, E. S., dos Santos Portela, C., Cordeiro, T. D., Bittencourt, I. I., Isotani, S., Melo, R. F., Muñoz, R., & Cechinel, C. (2024). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*.

Redyuk, S., Kaoudi, Z., Markl, V., & Schelter, S. (2021). Automating data quality validation for dynamic data ingestion [Published under the terms of the Creative Commons license CC-by-nc-nd 4.0]. *Proceedings of the 24th International Conference on Extending Database Technology (EDBT)*, March 23–26, 2021. https://openproceedings.org/2021/...

Restrepo-Carmona, J. A., Zuluaga, J. C., Velásquez, M., Zuluaga, C., Villamil, R. M., Morales, O., Hurtado, Á. M., Escobar, C. A., Sierra-Pérez, J., & Vásquez, R. E. (2024). Smart supervision of public expenditure: A review on data capture, storage, processing, and interoperability with a case study from colombia. *Information*, *15*(10), 616.

Rocha, J. C., Ramos, V., Cechinel, C., Hernández-Leal, E. J., Munoz, R., & Primo, T. T. (2024). Data interoperability in learning analytics - review of literature. *2024 L Latin American Computer Conference (CLEI)*, 1–8. DOI: https://doi.org/10.1109/CLEI64178.2024.10700464.

Sellar, S. (2015). Data infrastructure: A review of expanding accountability systems and large-scale assessments in education. *Discourse: Studies in the Cultural Politics of Education*, *36*(5), 765–777. DOI: https://doi.org/10.1080/01596306.2014.931117.

Srinivasan, K., Muthu, S., Devadasan, S., & Sugumaran, C. (2016). Six sigma through dmaic phases: A literature review. *International Journal of Productivity and Quality Management*, *17*(2), 236–257. DOI: https://doi.org/10.1504/IJPQM.2016.074462.

Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: A holistic approach to continuous quality management. *Journal of Big Data*, *8*(1), 76. DOI: https://doi.org/10.1186/s40537-021-00468-0.

Tavares, A. R., & Bitencourt, C. M. (2022). Avaliação de políticas públicas e interoperabilidade na perspectiva da governança pública digital. *Revista de Direito Econômico e Socioambiental*, *13*(3), 687–723.

Valdés, E., & Lecaros, J. A. (2024). Biobanks and data interoperability in latin america: Engendering high-quality evidence for the global research ecosystem. *Frontiers in Medicine*, *11*, 1481891.

Vasques, L., Rijo, R., & Alves, D. (2024). Interoperability at healthcare institutions in brazil: Framework for identify maturity stages. *Procedia Computer Science*, *239*, 2316–2321.

Wahid, A. M., Afuan, L., & Utomo, F. S. (2024). Enhancing collaboration data management through data warehouse design: Meeting ban-pt accreditation and kerma reporting requirements in higher education. *J. Tek. Inform. Jutif*, *5*(6).

Wijayanti, W., Hidayanto, A. N., Wilantika, N., Adawati, I. R., & Yudhoatmojo, S. B. (2018). Data quality assessment on higher education: A case study of institute of statistics. *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 231–236.

Williamson, B. (2019). Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education. *British Journal of Educational Technology*, *50*(6), 2794–2809.