Structuring Georeferenced Open Data of Brazilian Schools on Wikidata

Hector Gabriel Corrale de Matos
University of São Paulo

Rafael Ghidini
Santa Catarina State Department of Education

Túllio Morais Franca
Federal University of Goiás

Lilian Cássia Bornia Jacob
University of São Paulo

Abstract

This study demonstrates the integration of georeferenced Brazilian educational data into Wikidata, enhancing completeness and revealing regional disparities in open data on schools.

Keywords: Wikidata, Georeferenced, Education, Open Data, Brazil

Introduction

Structuring open data plays a critical role in education. It facilitates transparency, accessibility, and the interoperability of information. Integrating open repositories, particularly Wikidata, offers significant potential for curating information on educational institutions and promoting data literacy. Wikidata's organizational framework standardizes vast quantities of information for use by researchers, administrators, and policymakers. Georeferenced data is essential for mapping educational institutions' locations, identifying regional disparities, and informing evidence-based policies. Disseminating school data through Wikidata potentially promotes universal educational models while supporting data-driven initiatives and mechanisms for social participation.

INEP Schools Catalog and Wikidata: The main open data source for education in Brazil is provided by the National Institute of Educational Studies and Research Anísio Teixeira (INEP). The INEP Schools Catalog constitutes a comprehensive open dataset representing the Brazilian education system. It is hierarchically organized by educational level (early childhood, primary, secondary, and higher education), funding category (public or private), and territorial classification (rural or urban). Public education, which serves the majority of Brazil's population, is further geographically differentiated into municipal, state, and federal jurisdictions.

Schools Statistics: Brazil comprises 5,570 municipalities across 27 federative units, with a total of 178,476 schools according to the Basic Education School Census (INEP, 2023). Approximately 28.5% of these institutions are situated in rural areas, including 3,484 schools located within Indigenous territories. Socioeconomic disparities, ethnocultural factors, and geographical variations significantly influence the quality of educational access across Brazilian regions. These inequalities manifest in the availability of data and representational bias of Wikidata entries, as evidenced by the disproportionate concentration of information on schools in regions characterized by higher population density and economic development. Furthermore, there is little information about Indigenous communities, and including their schools in Wikidata can strengthen the visibility of these peoples by incorporating data such as the Indigenous languages taught and the traditional cultural and school events of each community.

Wikidata Landscape of School Data: As of March 2025, Wikidata contained 602,178 entries categorized as "schools" worldwide, with 8,741 of these located in Brazil. In this group, only 37% of the Brazilian school items contained the INEP code (Wikidata property P11704, a unique eight-digit identifier). However, although the INEP Schools Catalog is updated annually, some data may become outdated or incorrect. Therefore, using this dataset in Wikidata requires ongoing monitoring and, in some cases, verification against secondary databases and tools. This collaborative effort provides a complementary open-source data repository integrated with geolocation information. The enriched dataset offers valuable applications for resource allocation, policy planning, accessibility studies, and visualizing the distribution of educational infrastructure.

This research presents a mapping of information on educational institutions and documents the contributions of Brazilian editors in structuring open, georeferenced data on Wikidata. It details two primary interventions: the incorporation of INEP code identifiers into existing Wikidata entries and the creation of corresponding new Wikidata items for schools not yet represented. The study's objective was to integrate and validate data from the INEP Schools Catalog with existing Wikidata records, situating Wikidata as an additive and interoperable knowledge base.

Methods

This study is an exploratory-descriptive investigation utilizing secondary, publicly available data. All data sources are licensed under open terms, ensuring ethical compliance and avoiding exposure of sensitive information. The primary data source—the INEP Schools Catalog—was obtained in CSV format. This comprehensive dataset, derived from the annual School Census, contains information on all Brazilian educational institutions, including operational status; institutional names; INEP codes; geographical information; administrative classifications; contact details; partnership arrangements; regulatory frameworks; institutional dimensions; educational offerings; and precise geographic coordinates. Using this dataset, we carried out two major actions: (1) adding INEP identifiers to existing Wikidata items to guarantee interoperability between Wikidata and the School's Catalog, and (2) a pilot project to create new Wikidata items for schools missing from the database. The methodology was implemented between June and November 2024.

Interoperability: A comparative analysis between INEP records and Wikidata entries revealed significant discrepancies, necessitating a thorough reconciliation process to maintain data currency and integrity. The integration process proceeded through four structured phases: (i) Data acquisition – both the INEP Schools Catalog and relevant Wikidata data were downloaded as CSV files; (ii) Initial reconciliation – records were systematically compared in LibreOffice Calc using combinations of school name and municipality to group potential matches; (iii) Advanced matching – for records not automatically reconciled, the PHP function similar_text() was employed to compare school names and addresses, identifying high-confidence matches; (iv) Manual refinement – the remaining unreconciled cases underwent meticulous manual verification and adjustment. To aid ongoing data management, two worklists were created using the ListeriaBot tool: "Schools without INEP" (https://w.wiki/DF9Y) and "Schools by subtype without INEP" (https://w.wiki/DF9Z), which dynamically track school items missing the INEP identifier.

Creation of new items: A pilot study focusing on schools in Mato Grosso state (Central-West Brazil) served as a methodological tested, validating procedures for data integration, validation, and record updating. This pilot also assessed the availability of geographic coordinates and associated images for school records. Following reconciliation, all corresponding Wikidata entries were updated with the INEP code property (P11704). Data accuracy was evaluated by comparing the authoritative INEP Catalog information with the content of matching Wikidata items. SPARQL queries were used to analyze which school items had geographic coordinates and image illustrations (from Wikimedia Commons), providing metrics on data quality and completeness in Wikidata.

Results

The methodological framework successfully established interoperability between the INEP Schools Catalog and Wikidata, enabling seamless data integration. Analysis categorized the Brazilian school entries in Wikidata into two groups based on their semantic classification: those classified as an "instance of" a school and those designated as an "instance or subclass of" a school. In the first group, an initial data extraction (as of March 21, 2024) identified 3,506 school items in Wikidata, representing the national baseline prior to implementation of our methodology. Post-intervention (February 2025), Wikidata contained 6,430 school items in this category, of which 6,420 had INEP codes and 10 lacked this identifier. However, spatial and multimedia metadata remained limited; only 1,986 of these schools had geographic coordinates, and 110 had images. In the second group, representing schools tagged under a broader subclass hierarchy, Brazil's total was 2,176 school items (2,141 with INEP codes and 35 without). This corresponds to an INEP ID coverage rate of 98.4%, affirming the methodology's efficacy. Specifically, the Mato Grosso pilot initiative added 3,119 school items to Wikidata (Figure 1). Figure 1: Geographical context of Mato Grosso State in the Central-West region prior to the pilot initiative (upper panel), and the geospatial distribution of the 3,119 schools added in Mato Grosso following the initiative (lower panel), visualized as red dots. The upper map highlights the state's location in Brazil's Central-West region before implementation, while the lower map illustrates the post-intervention expansion of educational data, emphasizing the concentration of newly integrated schools in Mato Grosso. Source: https://w.wiki/DFJK

Conclusions

The analysis and implementation of our methodology for structuring open data on Brazilian schools in Wikidata focused primarily on data enrichment. By integrating and validating information from the INEP Schools Catalog with existing Wikidata records, this project serves as an example of effective data preservation and enhancement. The critical contribution lies in the assignment of INEP codes to school entries, which significantly improves data interoperability and creates meaningful connections between previously isolated datasets. This approach aligns with international and Wikimedia recommendations for fostering more equitable data representation, particularly regarding the Global South. Furthermore, the study emphasizes the need for community engagement and continuous dialogue with stakeholders to ensure that public data remain open and that such initiatives are sustainable. Representing educational data on Wikidata facilitates the visualization of critical issues for national educational policy and, through the creation of school items, strengthens the representation of communities and regions that have been notably underrepresented. These outcomes support knowledge equity through Wikidata's open knowledge framework.

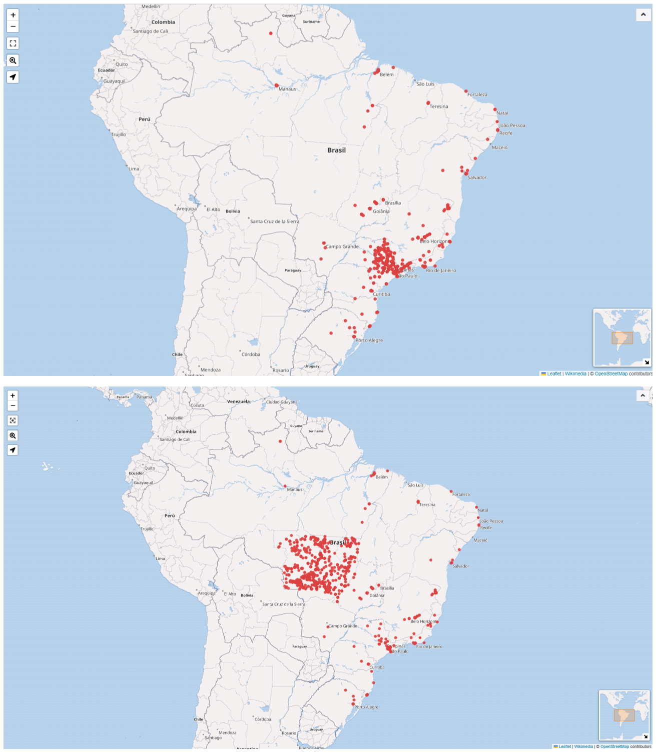

Figure 1: Geographical context of Mato Grosso State in the Central-West region prior to the pilot initiative (upper panel), and geospatial distribution of the 3,119 schools added in Mato Grosso following the initiative, visualized through red dots within the national context (lower panel). The upper map highlights the state's location in Brazil's Central-West region before implementation, while the lower map illustrates the post-intervention expansion of educational infrastructure, emphasizing the concentration of newly integrated schools in Mato Grosso. Source: https://w.wiki/DFJK

References

INEP. 2023. Basic Education School Census. National Institute of Educational Studies and Research Anísio Teixeira Open Data Repository. https://hdl.handle.net/20.500.14135/958.

IMPA. 2024. School Catalog. Institute of Pure and Applied Mathematics. https://buscainep.impa.br