Humanidata: semantic structuring of semantic academic data of humanities of Brazil and Latin America (1945-1969)

Erika Guetti Suca
Wikimedia Brasil

Ana Carla Rodrigues Ribeiro
State University of Maringá - Wikimedia Brasil

Isabela Tosta Ferreira
State University of Campinas - Wikimedia Brasil

Pedro T. Terres
State University of Campinas - Wikimedia Brasil

Abstract

This project addresses a critical gap in digital knowledge infrastructures by documenting Latin America's mid-20th century humanities scholarship through Wikidata. Between 1945 and 1969, academic conferences served as key hubs for intellectual exchange, yet their records remain scattered in physical archives or trapped in unstructured formats. Our work's goal is to systematically transform these materials into linked open data to reveal hidden scholarly networks. We employ mixed methods: (1) archival research to identify 150+ conference proceedings across five countries; (2) semantic modeling of participants, institutions, and themes using Wikidata's framework; and (3) computational analysis through SPARQL queries to map knowledge circulation patterns and gaps. The methodology innovates by bridging traditional historiography with open collaboration tools. Key outcomes include: (1) a dataset of Latin American humanities conferences; and (2) publicly reusable Wikidata tools that empower Wikimedia communities to expand content about pioneering Humanities scholars in Latin America. By making these academic networks visible and machine-readable, we challenge the marginalization of Global South scholarship in digital spaces. The project demonstrates how open knowledge platforms can recenter overlooked epistemic traditions, with direct applications for Wikipedia content creation and educational initiatives.

Keywords: Humanities Scholarly Data, Academic Profiles in Humanities, Web Semantic, Knowledge Gaps, Wikidata.

Introduction

Wikidata is at the heart of the Wikimedia projects, and is considered a central resource of the Wikimedia knowledge ecosystem. For many professionals and researchers, Wikidata is simply becoming the largest freely available knowledge graph today (Vrandečić et al., 2023). Yet, despite its global reach, critical gaps persist in the representation of Latin American academic and cultural history—particularly in the social sciences and humanities. Our proposal aims to address this gap by empowering Latin American Wikimedians with structured, reusable data on the region's intellectual networks, while also enhancing Wikidata's coverage of underrepresented scholarly traditions.

In the mid-20th century, Latin American countries already showed intense intellectual

activity through conferences and international collaborations. Yet much of this history exists only in fragmented archives, conference proceedings, or publications that lack structured metadata or doesn't even exist in digital spaces —making it difficult to integrate this knowledge into open projects, including Wikimedia projects, such as Wikidata. Without deliberate intervention, these intellectual networks risk remaining invisible, reinforcing the marginalization of Latin American scholarship in digital databases.

Our work aims to systematically extract and model data from historical academic conferences — participants, affiliations, funding sources, and thematic debates — into Wikidata. Most of this data exists only on physical media. By building this multilingual, open dataset of Latin American academic conferences (1945–1969), we intend to support community-driven research by developing SPARQL queries and visualization tools that help Wikimedians explore these networks, identify content gaps (e.g., women scholars missing from Wikipedia), and fuel community events for other Wikimedia projects.

Our interdisciplinary team combines historical research with technical expertise to achieve these goals. Two research assistants will conduct archival work, processing conference proceedings and academic materials to extract key metadata about participants, institutions, and funding sources. A PhD student will oversee the bibliographic research, ensuring rigorous contextualization of the data within regional academic traditions. Meanwhile, the project coordinator (Dr. Erika G. Suca, PhD in Computer Science) and a technical trainee will focus on the computational aspects—modeling the extracted data in Wikidata, developing SPARQL queries, and creating visualization tools for community use.

The long-term vision is one of infrastructural equity: ensuring that Latin America's knowledge ecosystems are not just consumers of Wikidata but active shapers of it. This project aligns with the Wikimedia 2030 Strategy's call to "close knowledge gaps" by providing the structured data needed to tell more complete stories — stories that, until now, have been constrained by the limitations of analog archives or proprietary databases. In doing so, we invite Latin American researchers, educators, and Wikimedians to engage with this data, challenge its narratives, and ultimately reclaim their place in the global knowledge commons.

Date: July 1, 2025 till June 30, 2026.

Related work

Existing scholarship has documented how transnational actors—particularly U.S. foundations (Ford, Rockefeller) and agencies (USAID, OCIAA)—shaped Latin American social sciences during the Cold War, prioritizing development agendas aligned with geopolitical interests (Miceli, 1993; Suprinyak & Fernández, 2022; Cunha, 2007). Parallel studies highlight the role of European institutions (e.g., French academic missions in Brazil) and UNESCO-backed networks (FLACSO, CEPAL) in concentrating intellectual capital while marginalizing peripheral actors (Abarzúa Cutroni, 2018; García, 2005; Soler, 2018). However, as Kreimer (2021) notes, these analyses remain fragmented by national frameworks or qualitative methods, failing to systematically trace the circulation of scholars, ideas, and funding across borders.

Our project addresses this gap by leveraging Wikidata's semantic architecture to map transnational academic networks, capturing relationships between scholars, institutions (e.g., CONICET, CNPq), and patrons (e.g., USAID, UNESCO). This approach bridges macro-level

policy studies and micro-histories of resistance, offering the first computational framework to analyze how epistemic power circulated—and was contested—in Cold War-era Latin America.

From a technical point of view, we mention some successful cases that explore the potential of Wikidata in the area of digital libraries and the representation of cultural heritage, historical data and academic publications from academic institutions. Evenstein and Nachmias (2023) discussed several successful case studies in promoting digital literacy through the use of Wikidata, demonstrating favorable results for the use of Wikidata by novice users. Among the different advantages found in the research, we highlight the creation of data visualizations, analysis and critical thinking, and the creation of improved data. Other skills developed include issues related to modeling, analysis, data validity and the verification of systematic data bias. Similarly, the work of Lemus-Rojas et al (2022) argues that Wikidata not only increases the visibility of scholars in an open environment, but also helps to contribute to improving the reputation of academic institutions. As scholarly data is added to Wikidata, researchers' work becomes easier to discover and cite.

Baker and Mahal (2024) analyzed Wikidata to create biographies of women archaeologists using Wikidata as a knowledge infrastructure. Their project, called Beyond Notability, created several biographies using linked data and allowed them to explore the ideologies that Wikidata adopts and reproduces to represent information in the biographies. However, the research pointed out many inattentions on the platform, especially in temporary aspects of people's activities, customs, cultural affiliations, and places frequented in people's history. The work allowed us to point out limitations in the Wikidata model and brought reflections on the features that need to be improved in the future of the semantic platform.

The research of Mihindukulasooriya et al. (2025) show that conference data is only minimally represented on Wikidata. They developed a methodology to use large language models (LLM) to efficiently extract conference metadata from multiple sources through a human validation process and using OpenRefine. They registered the scholarly data on Wikidata through QuickStatements. Their method is innovative, but it is viable since the information is already present on the internet and has already been made available in an open digital environment.

Zhao research (2022) points to Wikidata as helping with metadata enrichment and curation, data interoperability, and interconnection of linked data in knowledge modeling. However, it points to several problems for the Digital Humanities field, for example, technical problems due to lack of prior knowledge or training, as well as lack of participation from domain experts, or a clear and well-defined workflow that coordinates project tasks. Researchers Gutiérrez and Fontenelle (2025) also describe similar problems in their experience with Wikidata and Digital Humanities, highlighting the ontological confusion of professionals about what can be an instance of what, or the determination of a standard concept for experts, or the definition of when the model is complete.

The work of Suca et al. (2024) present a methodology that shows how relationships between academic data, metadata and other bibliographic identifiers can be enriched in Wikidata. And how, based on the semantically structured information in Wikidata, SPARQL queries can be performed that allow the implementation of personalized queries that enable a better understanding of the information. The research also emphasizes how

research analysis would be very difficult to obtain on other bibliographic database platforms due to the rigidity of the analysis possibilities with predetermined indicators or even the need to pay for the use of the platform. They developed a web page Vitrine NeuroMat that uses structured narrative models filled with information extracted from Wikidata. The page is automatically updated as soon as new information from the research group is made available on Wikidata. We adapted the methodology of this research to our work, especially in the technical phase of ingesting information in Wikidata.

Methods

This work is an exploratory research, with the main objective of enabling the visualization and analysis of the academic evolution of research in the humanities area from 1945 to 1969 mainly in the context of Brazil. Figure 1 presents the proposed methodology. We introduce our methodology in the following phases: (1) data collection, (2) structuring of information in shared spreadsheets, (3) mapping and semantic structuring with Wikidata metadata, (4) recording of information on Wikidata, finally, (5) analysis of information using the graphical potential of Wikidata.

The quality of the ontology developed depends on a good methodology. In view of this, we will apply good practices for ontology development following the literature (Kendall et al., 2019). Before starting data collection, we will raise basic questions, later specific to the domain, that will serve to understand the purpose of building the ontology. For example, what will be the uses of the ontology? Who will use and maintain the ontology? What types of questions should the ontology answer? These questions will be used both to establish a limit to the scope of the ontology and will serve as a basis for validating the ontology.

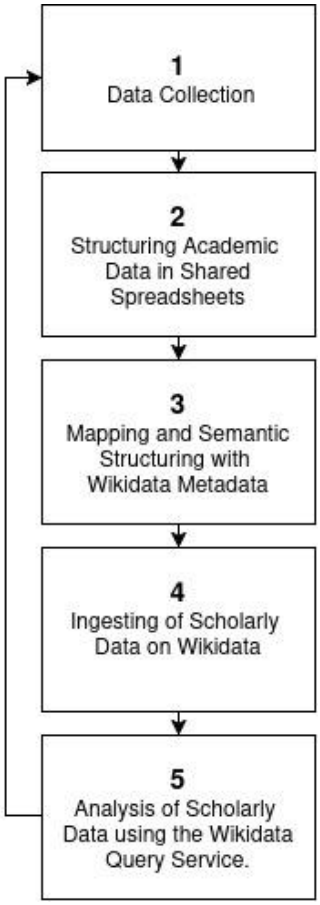

Figure 1. Methodology proposed to semantic structuring of semantic academic data of humanities of Brazil and Latin America from 1945 to 1969.

We will follow a spiral model of application of the previously proposed methodology. First, we will select the most relevant information to be recorded, such as: place, date, topics discussed, participants, etc. Then, in each spiral sequence, data validation and verification will be carried out, and then more events will be recorded, always linked to the events and information previously recorded.

The data collection process begins with a targeted acquiring of 150 conference proceedings ("anais" in Portuguese, "actas" in Spanish) between 1945 and 1969 from two key repositories for each studied country: Brazil, Argentina, Chile, Uruguay and Paraguay,

focusing on national and international events while excluding local/regional gatherings. These materials are prioritized because their structured records—containing participant lists, institutional affiliations, and thematic sessions—provide unique insights into intellectual networks often absent from traditional publications.

Searching the repositories using standardized keywords (e.g., "actas" + "congreso", "conferencia") and filtering for disciplines like history, philosophy, geography, anthropology and social sciences, while excluding STEM fields, we will create a list of conference proceedings that are within the scope of the project. With this list for each repository, we will contract local junior researchers to do the digitizing and pre-processing of materials.

With the proceeding digitized and pre-processed, the data pipeline then progresses through iterative phases: first, our team will structure information in shared spreadsheets with columns for core Wikidata properties related to conference proceedings (event location [P276], participants [P710], date [P585], sponsor [P859], etc.) From this *corpus*, we will select 50 representative conferences (10 per country) for full semantic enrichment and analysis. This curated subset will undergo rigorous cross-referencing, with particular attention to metadata enrichment for all participants - tracking institutional affiliations [P1416], educational backgrounds [P69], and gender [P21] where documented.
The final stages involve Wikidata integration using QuickStatements and analysis through customized SPARQL queries that visualize: 1) gender disparities in participation rates, 2) thematic evolution across years 3) geographic distributions of conferences by funding source (comparing UNESCO vs. Ford Foundation sponsorship patterns). This approach combines traditional historiographic rigor—through careful sourcing from Latin America's most comprehensive humanities collections—with computational scalability, enabling both macro-analysis of regional trends and micro-study of individual scholars' trajectories. The spiral model ensures continuous refinement, with each iteration incorporating new property relationships (e.g., "educated at [P69]" linkages) as the team's bilingual coordinators reconcile Portuguese/Spanish metadata variations.

Expected output

A primary scholarly output of this research will be a peer-reviewed scholarly publications, an article submitted to the *Journal of Digital History* (JDH), an open-access journal co-published by the Luxembourg Centre for Contemporary and Digital History (C^2DH), leveraging its unique three-layer format to bridge humanistic critique with computational reproducibility. Where traditional historical publications might separate methodological reflection from data presentation, JDH's layered structure allows us to simultaneously:

1. **Narrate** the epistemic challenges of reconstructing Latin American academia in the 1940s, 1950s and 1960s through Wikidata—questioning how semantic web technologies both reveal and obscure power dynamics in knowledge production. This topic aims to demonstrate how Wikidata - or other Wikimedia projects - can support quantitative and qualitative research in the Humanities;
2. **Demonstrate** our computational workflow with technical precision through Jupyter notebooks, featuring SPARQL queries that map collaboration networks and visualizations of institutional hierarchies (comparable to

CS conference papers in clarity). This topic aims to demonstrate technical solutions for data processing that make them more accessible and visualizable;

3. **Share** curated datasets as reusable, versioned assets in the data layer. This topic aims to demonstrate the global impact of contributing to knowledge as a service from the open source Wikidata database.

Other article will be submitted to journals such as *Digital Humanities Quarterly* (DHQ), *Journal of Cultural Analytics*, *Journal of Open Humanities Data* (JOHD), *Revista Latinoamericana de Tecnología Educativa* (RELATEC), *Revista Brasileira de Humanidades Digitais (RBHD)* and *Liinc em Revista* (UFRJ), depending on the flow of journals. We are going to use a CC-BY 4.0 for this article, to be able to make it compatible with Wikimedia projects. This publication would be specially targeted on digital humanities scholars, semantic web researchers, and Latin American studies specialists.

This project will generate multiple outputs designed to advance both scholarly understanding and public knowledge. The research will produce an intra-Wiki visualization dashboard hosted on Wikimedia Cloud Services, featuring customizable SPARQL queries that map gender disparities across three analytical dimensions: participation rates by country (comparing Brazil, Argentina, Chile, Uruguay and Paraguay), representation across conference themes (e.g., philosophy versus anthropology), and institutional affiliations. This living resource will automatically update as new conference data is added to Wikidata, serving as both a research tool and an educational exhibit about women's historical presence in Latin American humanities. All underlying datasets will be archived on Zenodo with DOI assignment, including: 1) cleaned conference proceedings with unified metadata standards, 2) participant lists enriched with gender and institutional markers, and 3) documentation of archival gaps to guide future research, and finally, 4) the project code will be documented and made available in a GitHub repository.

It is also intended that the research and its findings be shared at two digital humanities and scientific communication events in Latin America and in Wikimedia conferences such as Wikimania and other opportunities.

Risks

A risk analysis table was created with possibilities that may occur in the project, and possible actions to mitigate them. One of the main risk aspects of the project is the expectation of finding a larger amount of data than expected. As stated in the spreadsheet, the proposal in case this occurs is to work with filters and thematic cuts, such as the issue of gender, so that it is possible to conclude the research with good results. There are also risks related to interpersonal and work relationships, as can occur in any group. To mitigate this situation, we propose the creation of a welcoming environment, of active listening and based on the Wikimedia Brasil Friendly Space Policy. Here is the full risk analysis table.

Community impact plan

The project aligns with the Wikimedia Movement's 2030 strategy by delivering knowledge as a service and improving knowledge equity. Via our research, we aim to:

- Contribute directly to the inclusion of items on Brazilian and Latin American knowledge and culture in Wikidata;
- Create mapping of the properties of academic works, building a network of relationships between Latin American research centers;

- Contribute on the quality of data such as gender of researchers, region, research topic inclination by research centers that can be converted into maps;
- Consolidation of research from a perspective of knowledge produced by the global south.

All of this data can be converted into events such as webinars and scientific dissemination, which could be organized by the HumaniData Project in cooperation with other affiliates and user groups in Latin America. The objectives and execution of the research are aligned with the purposes of Latin American groups such as Wikimedia Argentina, which directly cite priorities in the areas of: local culture and decolonization, memory construction and digital culture, and open science. Thus, we believe that the data added to Wikidata can be used by these types of groups. Also, this data can possibly be used in projects that aim to consolidate the Wikimedia community at a Latin American and Brazilian level, mostly because Wikimedia projects such as Women in Red, Wiki Editoras Lx, the VisibleWikiWomen campaign, the Mais Teoria da História na Wiki project and others are using Wikidata queries as the source to create lists of non-existent Wikipedia articles on biographies and institutions to be edited in community events. In the same vein, campaigns such as Whose Knowledge's Visible Wiki Women would benefit on queries for women in the field of Humanities in Latin America that are lacking any images on Wikimedia Commons. Tools such as the "BHL Image Explorer", created recently for the Biodiversity Heritage Library GLAM, show how Wikidata may be used as an avenue to crowd-source contributions to both Wikimedia Commons and Wikipedia.

Evaluation

Success will be measured by our ability to deliver a robust, reusable dataset documenting Latin American humanities networks from 1940s to 1960s. The minimum viable outcome includes processing at least 150 conference proceedings through initial data extraction, with full semantic modeling of 50 key conferences (10 per country) in Wikidata. Each conference record will include structured metadata on participants, institutions, and themes, with particular attention to gender representation.

The project's academic contribution will be assessed through two primary outputs: a peer-reviewed article in the Journal of Digital History and presentations at both a computer science venue and a humanities conference. These publications will demonstrate how semantic web technologies can advance historical research. Beyond traditional academia, we will develop an intra-Wiki visualization dashboard enabling users to explore disparities across conferences, funding sources, and disciplines. Success here means not just creating the tool, but ensuring it is functional, documented, and actively used — with its data being used by at least one Wikimedia initiative from Latin America.

Budget

The budget includes the payment of researchers involved in the work, as well as the hiring of junior researchers to collect materials outside Brazil, where the researchers live. A budget was also considered for travel to conferences to share the work throughout its production. The budget is available here.

References

BAKER, J., MAHAL, A. "I have always found the whole area a minefield": Wikidata, historical lives, and knowledge infrastructure. **Int J Digit Humanities.** 2024. https://doi.org/10.1007/s42803-024-00090-5

BRUSCHI, Rita; CUTINELLA, César; PESCE, Fernando; BRUSCHI, Rita; CUTINELLA, César; PESCE, Fernando. Los orígenes de la comunidad geográfica en Uruguay. La Asociación de Geógrafos del Uruguay y la Revista Uruguaya de Geografía (1950- 1957). **Boletín de estudios geográficos**, Montevidéo, n. 120, p. 43–68, dez. 2023. https://doi.org/10.48162/rev.40.031

CUNHA, Luiz Antônio Constant Rodrigues da. **A universidade crítica**: o ensino superior na república populista. São Paulo]: UNESP, 2007.

CUTRONI, Anabella Abarzúa. El poder simbólico de la UNESCO en América Latina sobre el vínculo FLACSO - UNESCO. **História da Educação**, São Paulo, v. 22, p. 244–259, ago. 2018. https://doi.org/10.1590/2236-3459/72083.

EVENSTEIN Sigalov, S., NACHMIAS, R. Investigating the potential of the semantic web for education: Exploring Wikidata as a learning platform. **Educ Inf Technol 28**, 12565–12614. 2023. https://doi.org/10.1007/s10639-023-11664-1

GARCIA, Afrânio. Circulation internationale et formation d'une "ecole de pensee" latino-americaine (1945-2000). **Social Science Information**, v. 44, n. 2–3, p. 521–555, 1 jun. 2005. https://doi.org/10.1177/0539018405053296.

GUTIÉRREZ, Silvia; Fontenelle, Giovanna. Main Challenges of Wikidata for Librarians. In: Diff, the Wikimedia Blog. https://diff.wikimedia.org/2024/06/27/ld42023-v-main-challenges-of-wikidata-for-librarians/ Accessed: 15 April 2025.

KENADLL, F. Elisa; McGuinness, Deborah L., and Ying Ding. 2019. Ontology Engineering. Morgan & Claypool Publishers.

KREIMER, Pablo. **Science and Society in Latin America: Peripheral Modernities**. New York: Routledge, 2019.

LEMUS-ROJAS, M., Odell, J., Brys, L. & Ramirez Rojas, M. Leveraging Wikidata to Build Scholarly Profiles as Service. In: **Knowledge Creation, Dissemination, and Preservation Studies (KULA)**, 6(3), 1–14. 2022. https://doi.org/10.18357/kula.171

LORENA, Soler. **Los oficios del Sociólogo en Paraguay (1950-1980)**. Buenos Aires: FLACSO, 2019. Acesso em: 4 abr. 2025.

MANSKE, M. **QuickStatements**. Toolforge, 2014. Available at: https://quickstatements.toolforge.org/. Accessed: 10 April 2025.

MICELI , Sergio. **A Fundação Ford no Brasil**. São Paulo: Sumaré/Fapesp, 1993.

MIHINDUKULASOORIYA, N., Tiwari, S., Dobriy, D., Nielsen, F.Å., Chhetri, T.R., Polleres, A. Scholarly Wikidata: Population and Exploration of Conference Data in Wikidata Using LLMs. In: Alam, M., Rospocher, M., van Erp, M., Hollink, L., Gesese, G.A. (eds) **Knowledge Engineering and Knowledge Management**. EKAW 2024. Lecture Notes in Computer Science(), vol 15370.

Springer, Cham. 2025. https://doi.org/10.1007/978-3-031-77792-9_15

NEUROMAT **Vitrine**, 2025. Available at: https://vitrine.numec.prp.usp.br Accessed: 10 April 2025.

NICODEMO, Thiago Lima; SANTOS, Pedro Afonso Cristovão dos; PEREIRA, Mateus Henrique de Faria. **Uma introdução à história da historiografia brasileira (1870-1970)**. Rio de Janeiro: FGV Editora, 2018.

QUEIJO OLANO, Juan Andrés; JUNG, María Eugenia. Ideas y proyectos sobre la modernización de la Universidad en Uruguay (1950-1985). **Nuevo Mundo Mundos Nuevos**, Paris, 21 jun. 2022. DOI 10.4000/nuevomundo.88446. Disponível em: https://journals.openedition.org/nuevomundo/88446. Acesso em: 4 abr. 2025.

SUCA, Erika Guetti; BURLEY, Paul R.; ALVES, Éder Porto Ferreira; PESCHANSKI, João Alexandre. WIKIDATA for Structuring Academic Institutional Profiles: a NEUROMAT scholarly publications case study. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, *[S. l.]*, v. 30, p. 1–23, 2024. DOI: 10.5007/1518-2924.2025.e100592. In: https://periodicos.ufsc.br/index.php/eb/article/view/100592.

TERRES, Pedro Toniazzo; PIANTÁ, Lucas Tubino. Wikipédia: públicos globais, histórias digitais. **Esboços: histórias em contextos globais**, Florianópolis, v. 27, n. 45, p. 264–285, 19 jun. 2020. https://doi.org/10.5007/2175-7976.2020.e68391.

VRANDEČIĆ, Denny , Pintscher, Lydia and Krötzsch, Markus. 2023. Wikidata: The Making Of. In **Companion Proceedings of the ACM Web Conference 2023** (WWW '23 Companion), April 30--May 04, 2023, Austin, TX, USA. ACM, New York, NY, USA 10 Pages. https://doi.org/10.1145/3543873.3585579

ZHAO, Fudie, A systematic review of Wikidata in Digital Humanities projects. In: **Digital Scholarship in the Humanities**, Volume 38, Issue 2, June 2023, Pages 852–874, https://doi.org/10.1093/llc/fqac083