Social Memory about Chileans in Wikipedia: Politicians, Scientists, Artists, and Sportspersons since the 19th century.

Pablo Beytía (1, 2, 3, 4) Carlos Cruz Infante (1, 5)

(1) Monitor Social, (2) Catholic University of Chile, (3) Humboldt University of Berlin, (4) Weizenbaum Institute, (5) Università di Roma La Sapienza

Abstract

We propose investigating Chilean biographies on Wikipedia, examining how content gaps evolve across generations of notable people. Our goal is to explore these gaps in four occupational domains (politics, science, art, and sport) and understand, with these trends, how Wikipedia is framing (highlighting aspects of) the history of notable Chileans. Our research methods include (1) data extraction from Wikidata, (2) a manual database review process, and (3) natural language processing (named-entity recognition). We will deliver four onputs: (1) a database with biographies of Chileans and their essential characteristics, (2) a database with entity recognition analysis, (3) open-access documentation to extract the data and create similar projects, (4) a report of results presented in the wiki Workshop 2024, in Wikimedia Chile, and the local community of Wikipedians. As a next step, we expect to publish our analyses in an academic journal and create interactive tools that allow Wikipedians to explore their biases or blind spots easily.

Introduction

Wikipedia is an inclusive project, as it seeks to provide people with access to all types of knowledge in an unbiased manner (Wikipedia, 2023). Its official purpose is to be *accessible* ("a widely accessible and free encyclopedia"), *diverse* (a compendium of "information on all branches of knowledge"), and *impartial* (including "neutrally written" articles) (Wikipedia, 2023).

However, providing diverse and impartial knowledge is challenging. To the best of our knowledge, Wikipedia has multiple "knowledge gaps," and several are related to social group representation (Redi et al., 2021). For instance, several biases have been found regarding gender, sexual orientation, birthplace, and the historical era wherein the biographed people lived (Beytía, 2020; Redi et al., 2021; Reznik & Shatalov, 2016; Wagner et al., 2016).

Against this backdrop, we pretend to understand and illustrate how Wikipedia's biographies storage "frames" (Entman, 1993, 2007) the history of notable persons in a specific country, namely Chile. To do so, we will analyze the information patterns of Chilean notables' biographies, considering variables such as gender, place of birth, occupation, and the diffusion of biographies in multiple languages. We will examine content asymmetries in these and other variables by looking at the evolution of content imbalances by the birth generation of the notables.

In other words, we seek to know how Wikipedia selects aspects of the perceived reality (the Chileans' history) to make them more salient in communication.
The following questions will guide our research:

- How do Wikipedia's "knowledge gaps" (in dimensions such as gender or birthplace) evolve over generations of notable Chilean individuals?
- How does the historical trajectory of content asymmetries diverge in different occupational dimensions (politics, science, art, and sport)?
- How do these trends "frame" (highlight aspects of) the history of notable Chileans, as reported by Wikipedia?

We expect to identify the main asymmetries of social representation that Wikipedia inadvertently introduces when documenting the history of notable Chileans. We believe that this could help Wikipedians to become more aware of their biases and blind spots, and thereby advance in providing a more diverse and unbiased encyclopedic record of biographies. We also consider that the completion of this project could potentially benefit the Wikipedia research community, since our first priority will be to generate documentation on all the data extraction and analysis pipelines we use in this research. That would serve to easily replicate our analyses in other contexts (different countries, occupations, and historical periods).

Research period:
Start: July 1, 2023.
End: June 30, 2024.

Related work

A myriad of studies have shown that the biographies on Wikipedia tend to make visible specific people categories. For example, they are significantly focused on men (Hinnosaar, 2019) born in the Global North (Beytía, 2020), who lived in the last century (Samoilenko et al., 2017), and who excelled in professions such as mass arts or popular sports (Reznik & Shatalov, 2016).
We have investigated biographical content asymmetries related to gender and place of birth:

1. In a study based on the Networked Pantheon dataset (Beytía & Schobin, 2020), we found that only five countries in the Global North concentrate more than 62% of Wikipedia's biographical coverage. In addition, we estimate that the inequality in coverage between countries reaches a Gini coefficient of .84 (Beytía, 2020).
2. We examined the written and visual asymmetries between men's and women's biographies in the ten most widely spoken languages (Beytía et al., 2022). We concluded that (a) most of the male bias arises when selecting who will have a biography, (b) written and visual asymmetries do not follow the same patterns of disparities, (c) men biographies tend to have more images across languages, and (d) female biographies average better visual quality.
3. In another study, we proposed a general theoretical framework to closely observe content asymmetries in Wikipedia, which was tested with research findings on gender gaps (Beytía & Wagner, 2022). Our "Visibility Layers" model serves to analyze content inequalities across three editorial stages (content selection, building, and positioning) that contribute to making groups of biographies more or less visible.

The literature mentioned above does not analyze knowledge asymmetries from the historical account point of view. A second line of research has analyzed temporal evolution by looking at specific variables. For instance, biographies have been examined to observe the evolution of occupations across generations (Jara-Figueroa et al., 2019), variation in

migration patterns (Menini et al., 2017), and changes in biographical ties (Schich et al., 2014). These studies analyze content-specific variables (occupational distribution, migration, biographical relationships) and not the joint evolution of *multiple* content asymmetries across generations. Therefore, they do not offer a comprehensive look at how biographies frame the history of any specific human group.

In a recent study, we attempted to develop a more comprehensive approach, examining how Wikipedia frames the history of sociology in multiple and interrelated information structures (Beytía & Müller, 2022). There, we studied content structures and asymmetries across all generations of notable sociologists.

Through the current research proposal, we aim to replicate the "Wikipedia-framing" approach, albeit with a different subject matter. Instead of studying the history of a specific scientific discipline, we propose investigating how Wikipedia frames the history of notable people in a specific country and across different professions. To our knowledge, this topic has not been investigated hitherto now, especially examining the joint analysis of multiple content asymmetries that vary across generations.

Methods

We will focus on specific occupational categories that represent socially diverse areas of notable people: politicians, scientists, artists, and sportspersons. Previous research has shown that, in the last generations of notable people (those born since the 1950s), those occupational dimensions are the ones with the highest number of biographies on Wikipedia (Reznik & Shatalov, 2016; Yu et al., 2016). Except for sportspersons, those same categories are also the dominant occupations in the generations born since the beginning of Chile's history (early 19th century).

We will analyze the biographies of those occupational dimensions using a five-stage methodology:

1. *Data extraction*: we will create databases of people with Chilean nationality in the four occupational domains using Wikidata. From that source, we will extract essential biographical information (name, gender, year of birth, birthplace, year of death, death place, a portrait, and biography hyperlink). We will also obtain information on the notables' participation in sub-domains (i.e., political parties, scientific disciplines, sports branches, and artistic disciplines). As an approximation to the communicative influence of each notable, we will get the number of available languages for each biography.
2. *Preprocessing*: we will manually check the data for correctness. We will also complete the record if any relevant biography is missing and available on Wikipedia.
3. *Natural language processing (NLP)*: we will perform NLP on the Spanish biographies to recognize relevant discursive entities (people, events, places, dates, and organizations). We will start our analysis from the hyperlinks extracted from Wikidata. Then, we will extract the biographical text and employ the Python library Spacy to develop a named-entity recognition (NER). We will do this processing separately for each occupational domain. Finally, we will review the lists of recognized entities (people, events, places, dates, and organizations) to check for repeated terms, incomplete names, or nonsense phrases.
4. *Research products*: we will prepare (1) an open database of notable Chileans in Wikipedia, (2) an open database of the main entities that we found in the NLP, and (3) scripts and documentation on how to extract this data from Wikidata and perform the named-entity recognition

process with Wikipedia biographies. All this material will be released on an open-access project archiving platform (such as SocArXiv) under CC BY-SA 4.0 or a more permissive license.

Due to budget limitations, we will not be able to have a published scientific article in this research period. However, we will send an article proposal to the Wiki Workshop 2024, to present our work in progress. Our presentation will show how Wikipedia frames the history of notable Chileans and graphically depict the content asymmetries observed with our data. We will especially examine the asymmetries of gender, place of birth, and occupational sub-domains. This report will show whether these asymmetries tend to vary together or separately across generations and occupational domains. It will also identify generations of notable Chileans who produce relevant content breaks. (The article by Beytía & Müller (2022) is a good reference for the type of results that we want to show in this report).
To facilitate the use of our data by Wikipedians, we will aim to develop all graphics presented in the Wiki Workshop as interactive web tools. That will allow Wikipedians to explore specific patterns and biases on their own. This fund will not finance this research stage, but it is an internal goal of our social analytics project (www.monitorsocial.cl). As a reference for these interactive tools, we have already piloted a dashboard on Chilean politicians (https://www.monitorsocial.cl/politicos).
The following figures show some of the tools we have developed to analyze biographical information in Wikipedia.

Figure 1. Politicians' cards by political party

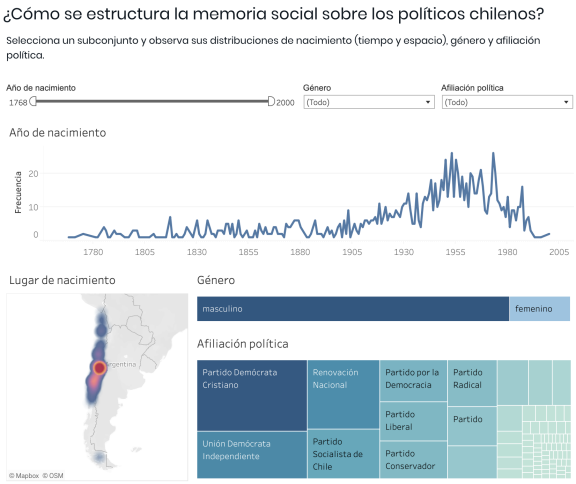

Figure 2. Content asymmetries explorer

Figure 3. Biographies search engine

Expected output

The research outputs will be the following:

- *An open-access database of notable Chileans in Wikipedia*, focused on remarkable politicians, scientists, artists, and sportspersons. That would be useful for researchers interested in knowledge gaps on Wikipedia and Wikipedians trying to bridge the content gaps in the Chileans' biographies.
- *An open-access database of the most relevant entities found in the biographies text*. This database includes lists of the most mentioned persons, places, and institutions in the biographies of Chilean notables. This input would be useful for researchers interested in Wikipedia's discourse and Wikipedians trying to understand the biases present in the Chileans' biographies' text.
- *Open-access documentation and scripts to extract our data from Wikidata and perform the named-entity recognition process with Spacey*. That might be useful input for those interested in replicating the Chilean case or the same research for other countries and occupations.
- *A presentation in Wiki Workshop 2024,* showing our main results (the work in progress to make a scientific publication). This will be useful for the Wikipedia research community, which will look at the results of the study and know where to find the documentation to replicate it.
- *A presentation of our results to Wikimedia Chile and the Wikipedians community in Chile*. This will directly help the volunteer editors who are trying to generate a more diverse and unbiased record of the Chileans' biographies.

Risks

Given the budgetary constraints, this research does not contemplate full dedication to writing a scientific paper. Instead, its priority is developing the databases and documentation so that the community of Wikipedians in Chile has updated information on the biographies' content biases and can easily replicate, improve or extend our research. The priority of data and documentation implies the risk that our analysis will not be formally published to a scientific audience.

We take the following precautions so that the analysis can be used to its fullest potential by volunteers and scientific audiences:

1. We will present the results on Wikimedia Chile and, through them, to the local community of Wikipedians.
2. We will make a work-in-progress report at the Wiki Workshop 2024, with the intention of ensuring the presentation of results to the scientific community and advancing a future publication.
3. When we prepare our presentation of results for Wikimedia Chile and the Wiki Workshop, we will create some interactive web tools to analyze the data. This will ensure that our analysis is open to exploration by the community.

Community impact plan

We have already contacted Wikimedia Chile so that they can advise us in some stages and plan together the dissemination of the results. This project connects with the 2030 Wikimedia Strategic Direction in multiple ways:

Knowledge as service:

1. It uses Wikidata's infrastructure to provide content that makes sense to a community and allows learning about it.

2. It encourages using Wikipedia knowledge to understand countries' history and memory.
3. It serves the community of Wikipedians in Chile, highlighting those social groups that do not have sufficient information and could be targeted for editorial work.

Knowledge equity:

1. It contributes to clarifying the groups excluded from Chilean politics, sciences, arts, and sports.

Plans for dissemination:

A Communications Officer will disseminate the content in three main ways:
Our media:

- Posts on our Website and Twitter account –over 10K followers.

Partnerships:

- Joint dissemination with Wikimedia Chile (we could organize a launch with Wikipedians).
- Academic dissemination through our networks and their social media accounts.
- An open-access webinar in partnership with *Punto Equality*, a Chilean NGO, presided by the former Minister of Women and the Gender Equality of Chile, Isabel Plá (to confirm).

Advertising:

- Social media paid advertising for disseminating the most impactful results.

Evaluation

An appropriate evaluation of this project might include the following questions:

1. Were the databases published on an open-access platform?
2. Were the codes and documentation necessary to create a database similar to this one published on an open-access platform?
3. Was the data analysis presented at the Wiki Workshop 2024?
4. Do the Wiki Workshop presentation answer the scientific questions of the project?
5. Were the results of this project presented in Wikimedia Chile?
6. Were the data and reports disseminated to the local Wikipedian community?
7. Is an effective plan applied to communicate these results on social media and to relevant local stakeholders?
8. How did the communication activities impact traffic, interactions, and downloads?

Budget

Our budget primarily considers the work of a group of researchers who will create the database (with the necessary scripts and documentation), analyze the data, prepare de public presentations, and communicate the results in collaboration with Wikimedia Chile.

The summary of the budget is the following:
- Salary or stipend: $13.000
- Communications and advertising: $1.900
- Benefits: $600
- Institutional overhead: $2.500

References

Beytía, P. (2020). The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia. Companion Proceedings of the Web Conference 2020, 806–810.

Beytía, P., Agarwal, P., Redi, M., & Singh, V. (2022). Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages. AAAI Conference on Web and Social Media (ICWSM).

Beytía, P., & Müller, H.-P. (2022). Towards a Digital Reflexive Sociology: Using Wikipedia's Biographical Repository as a Reflexive Tool. Poetics, 101732.

Beytía, P., & Schobin, J. (2020). Networked Pantheon: A Relational Database of Globally Famous People. Research Data Journal for the Humanities and Social Sciences, 5, 1–16. https://doi.org/10.1163/24523666-00501002

Beytía, P., & Wagner, C. (2022). Visibility Layers: A Framework for Facing the Complexity of the Gender Gap in Wikipedia Content. Internet Policy Review. https://doi.org/10.31235/osf.io/5ndkm

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication, 43(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Entman, R. M. (2007). Framing bias: Media in the distribution of power. Journal of Communication, 57(1), 163–173.

Hinnosaar, M. (2019). Gender inequality in new media: Evidence from Wikipedia. Journal of Economic Behavior & Organization, 163, 262–276.

Jara-Figueroa, C., Yu, A. Z., & Hidalgo, C. A. (2019). How the medium shapes the message: Printing and the rise of the arts and sciences. PloS One, 14(2), e0205771.

Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., & Lepri, B. (2017). Ramble on: Tracing movements of popular historical figures. Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 77–80.

Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2021). A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). ArXiv:2008.12314 [Cs]. http://arxiv.org/abs/2008.12314

Reznik, I., & Shatalov, V. (2016). Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. Journal of Informetrics, 10(1), 124–131.

Samoilenko, A., Lemmerich, F., Weller, K., Zens, M., & Strohmaier, M. (2017). Analysing timelines of national histories across wikipedia editions: A comparative computational approach. Eleventh International AAAI Conference on Web and Social Media.

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. Science, 345(6196), 558–562.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in Wikipedia. EPJ Data Science.

Wikipedia. (2023). Wikipedia:Purpose. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Purpose&oldid=1144512635

Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific Data, 3, 150075.