SOCIAL MEMORY ABOUT PEOPLE FROM A COUNTRY. THE CASE OF NOTABLE CHILEANS IN WIKIPEDIA

Pablo Beytía
Catholic University of Chile

Camila Rojas
Catholic University of Chile

Carlos Cruz
Sapienza University of Rome

Abstract

Wikipedia biographies are an influential form of social memory about human beings. This article describes a systematic method to explore the structure of this memory in biographies of people with the same nationality. The results are exemplified by analyzing the biographies of notable Chileans across generations.

Keywords: social memory, digital discourse, knowledge gaps, biographies, national history

Introduction

Wikipedia is an extraordinary social memory device about human lives. In its more than 320 languages, it includes biographies of over 6.2 million people identified on the platform as "notable" (Beytía et al., 2022), as they have received significant coverage from reliable sources that are independent of those individuals.

How this information is organized is highly influential. Wikipedia receives over 4 billion visits monthly and is the most visited online reference material globally (SimilarWeb, 2024). Large technology companies also use its information to define facts and train artificial intelligence models (Wang et al., 2019; The Economist, 2021).

It is important to examine in detail how the information propagated on Wikipedia is structured, as it significantly influences the global organization of knowledge and the production of cultural trends worldwide.

This research, funded by the Wikimedia Research Fund, presents a method for systematically analyzing social memory about people and generations from specific countries on Wikipedia. To exemplify the method and its possible results, it analyzes the information patterns of biographies of notable Chileans.

Related work

Studies have shown that Wikipedia has knowledge gaps (Redi et al., 2020) and tends to offer biographies of specific categories of people, particularly men (Hinnosaar, 2019), individuals born in the Global North (Beytía, 2020), and persons who excelled in professions such as mass arts or popular sports (Reznik & Shatalov, 2016).

Some studies have explored information structures from a historical or generational perspective, assessing how the biographical information structure changes according to the life span of notable individuals. For example, the birthplace distribution of notable individuals in different epochs (Schich et al., 2014) and the occupation distribution across generations (Jara-Figueroa et al., 2019) has been investigated.

The literature has focused on analyzing the complete record of biographies without exploring social memory about specific collectives (e.g., a nation). It has also specialized in analyzing particular gaps in the people portrayed (e.g., gender biases) rather than examining the biographical record in a multidimensional way, where these gaps vary in a coordinated manner

In contrast, our objective is to present a method to analyze *multidimensionally* the biographies of a specific collective (a nation) across generations.

Methods

Our approach attempts to understand the social memory and discourse of Wikipedia biographies as a phenomenon composed of several interrelated dimensions (see, for example, Beytía and Müller 2022):

- *Spatiality*: biographies represent people born and lived in certain territories.
- *Temporality*: they were born and lived in specific periods.
- *Attributes*: they possess multiple analyzable characteristics, such as their gender, occupation, and membership in institutions.
- *Semantics*: their biographies include text, images, and other multimodal symbols that create meaning about these

human lives.
- *Associativity*: biographies connect and form networks through hyperlinks and mentions, generating discursive clusters and differences in content centrality.
- *Multilinguality*: biographies vary by their number of language versions and language-specific configuration.

Our methods consider all these aspects of social memory organization. We evaluated them with biographies of notable Chileans. The computational processes are detailed on the project's official website. The most important steps are:

1. *Data mining*: we used Wikidata to extract basic information on all Chilean or Chilean-born notables (N=9,309). The information includes name, gender, birthdate, birthplace, death date, death place, portrait, occupations, biography hyperlink, and language versions.
2. *Occupation classification*: the notables had 699 occupations. Following previous studies, we distinguished 13 broad occupational categories. Then, we used GPT-3.5 to automatically classify occupations into them and evaluate reliability. We then manually verified the result.
3. *Natural language processing*: we extracted the articles' text and applied Named Entity Recognition (NER). We used the Flair Python library to identify people, organizations, places, and events mentioned in the biographies (F1-Score of 90.54 on the Spanish dataset).
4. *Network analysis*: we extracted the hyperlinks between the biographies to analyze the network of biographical associations in the occupational categories.
5. *Data visualization*: we used several visualization software (such as Plotly, Flourish, Gephi, and Looker Studio) to create graphs and interactive tools for visualizing the structure and semantics of biographies.

Results

This research is still ongoing. Its website includes open databases, interactive graphs, and a dashboard to interactively analyze the results. Some preliminary results about the social memory of notable Chileans are shown in Figures 1 and 2. There we identify some general trends:

- The record is significantly concentrated on notables of the 20th century (Figure 1A).
- There is consistently more memory about notable men, although women's record has grown proportionally since the generations born at the end of the 19th century (Figure 1B).
- At the beginning of the Chilean Republic (19th century), politicians, military men, and lawyers were the most visible figures, while sportsmen and artists have stood out in recent decades (Figure 1C).
- Throughout the generations, Chileans' average levels of diffusion on multiple languages are fairly stable (Figure 1D).
- Biographies about people from one occupational dimension tend to highlight the relevance of people from that same occupational dimension (Figure 2).

There are also examples of the network of hyperlinks between biographies (Figure 1E) and geographic trends in births adjusted for gender (Figure 1F).

Documentation for replicating this study in different notables and countries is open and available on the project website.

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*.

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.

Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2021). A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). ArXiv: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.

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.

Similarweb. (2023). *wikipedia.org. Análisis de tráfico y cuota de mercado*. https:// www.similarweb.com

The Economist. (2021, January 7). *The other tech giant—Wikipedia is 20, and its reputation has never been higher*. Archive. Ph. https://archive.ph/ZJkp6

Wang, C., Li, M., & Smola, A. J. (2019). *Language Models with Transformers* (arXiv:1904.09408). arXiv.

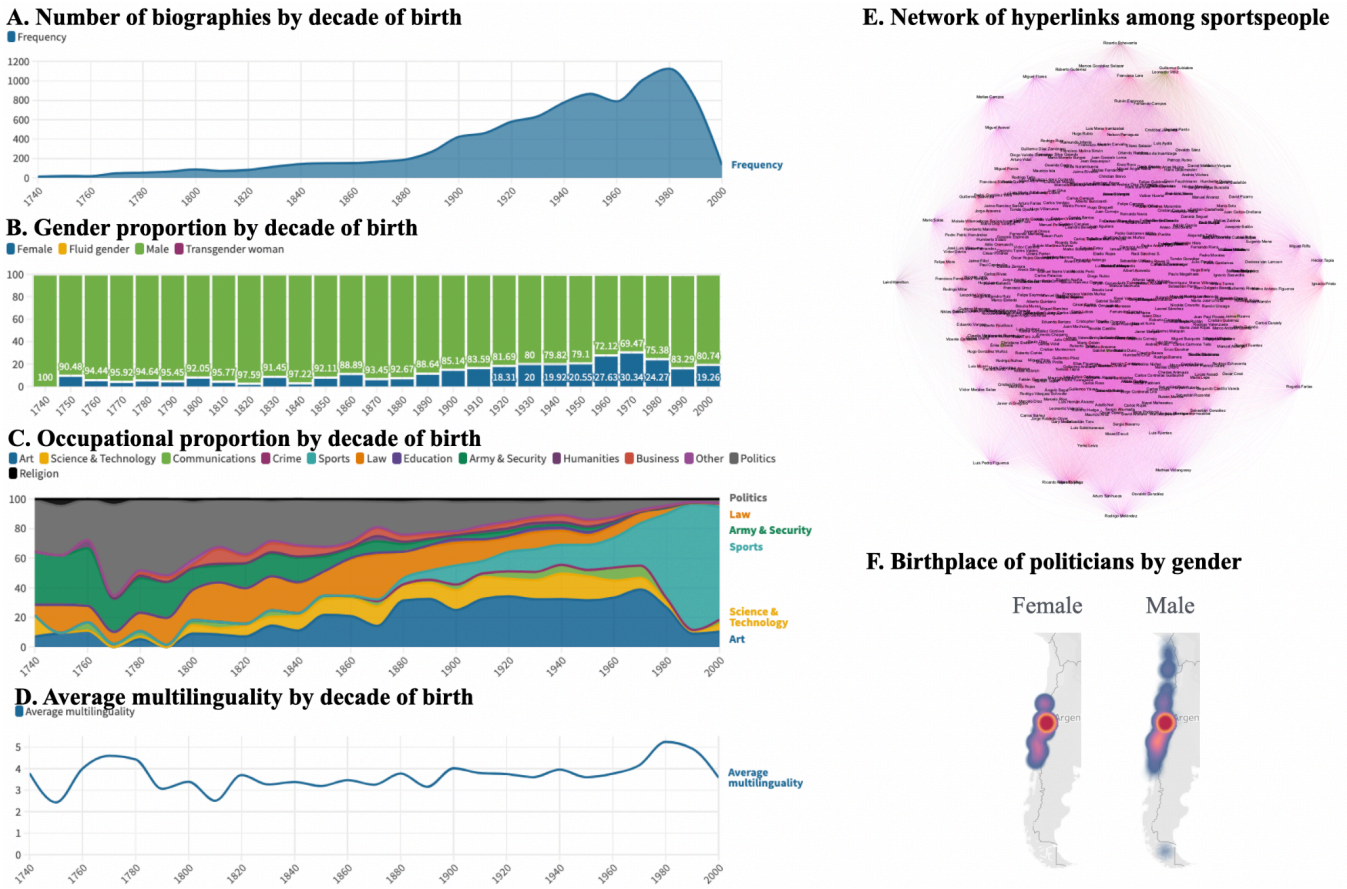

Figure 1: Biographical structures of notable Chileans

Figure 2: Named Entity Recognition in biographies of Chileans: most mentioned people in four occupational categories.