International Cartographic Association
Association Cartographique Internationale

The semantic street map guide: research on people that name our streets and what it says about our society

Maria Cristina Calvo Guinea [a,*], Alicia González Jiménez [a]

[a] *Spanish Geographic National Institute – mccalvo@transportes.gob.es , agjimenez@transportes.gob.es*

* Corresponding author

Abstract: The street network in urban settlements reflects decades—or even centuries—of history and transformation. Street layouts, zoning from gradual expansion, and architecture all offer insights into a city's development. But beyond physical form, street names also reveal the history and traditions of the area, exposing past power dynamics and the political views of those who chose them.

Often, the individuals behind these names are largely unfamiliar to the public. This lack of recognition makes it harder to perceive the inequalities embedded in this dimension of the urban landscape. Yet, even without knowing their specific identities, certain patterns are readily apparent—for instance, the overwhelming presence of male figures compared to female ones, among those commemorated in our street names.

Many projects analyse urban toponymy from historical, social, or spatial perspectives, often focusing on individual cities through detailed case studies. However, examining larger datasets at regional or national levels could offer deeper insights into societal evolution. Such analysis requires automated methods to reliably identify the individuals behind street names and collect relevant information about them.

This article outlines the procedures developed by the Spanish National Geographic Institute to characterize the street network according to the identities of the individuals after whom streets are named in the Madrid region, which is home to over 7 million people across 180 municipalities.

Keywords: streets, history, toponymy, settlements, Linked Data

1. Introduction

As Villar (2022) notes, the naming of urban streets and spaces is deeply tied to customs, traditions, and the political use of collective memory. These names form part of a shared identity that deserves to be studied and preserved. Moreover, they serve as a valuable resource for understanding the social, economic, and political contexts of different historical periods.

A clear example of this is the unequal distribution of street names in Spain, with significantly more named after men than women—highlighting the historically limited role of women in public life until recent decades. Another notable case is the persistence of streets named after individuals associated with the dictatorship that ruled from 1939 to 1975, despite the Democratic Memory Law, which requires local governments to rename streets honouring those accused of crimes during that period. However, there are currently no official statistics tracking this gender disparity or monitoring compliance with the renaming provisions of the law.

In recent years, various studies have explored the gender gap in street naming within specific medium- and large-sized cities. Most of these have emerged from volunteered geographic information initiatives or personal research efforts lacking long-term institutional support. The methodologies typically rely on manual data collection through online searches to gather relevant information.

Given the large volume of data involved, a manual, case-by-case approach is unfeasible for large-scale analyses. Spain has over 8,000 municipalities, most containing multiple population settlements. The authority to change street names lies with each municipality's Town Council, and any changes must be reported to the Spanish National Statistics Institute (INE). The INE publishes an updated list of street names every six months, which currently includes over one million entries. Clearly, analysing such a vast dataset requires automated methods for data collection and classification.

This article aims to present the methodology developed for large-scale analysis and classification of street names, along with the results obtained from its application in the Community of Madrid.

2. Methodology

2.1 Datasets

The analysis was conducted using the street network published by the Spanish National Geographic Institute. This dataset integrates official street names and codes from

the Spanish National Statistics Institute, along with geospatial data from various sources. In this case, the primary geospatial source was the official street map and gazetteer developed and continuously updated by the Regional Statistics Institute of the Community of Madrid.

for geospatial data. It represents the network at a 1:5.000 - 1:25.000 scale.

2.2 Identification of streets named after people: Named Entity Recognition

The dataset included approximately 45,000 street names. The first step, therefore, was to identify and extract those likely referring to individuals.

This was achieved using Natural Language Processing techniques, specifically the Named Entity Recognition (NER) model "roberta-base-bne-capitel-ner." This model is a fine-tuned version of the roberta-base-bne model, which was pre-trained on the largest Spanish corpus available to date—570GB of clean text, compiled from web crawls conducted by the National Library of Spain between 2009 and 2019. The text was specifically processed to serve as input for NER models. The process of creating these models is explained in detail in Gutiérrez-Fandiño et al. (2022).

From the original dataset, nearly 13,000 entries were identified as referring to individuals, resulting in approximately 7,000 unique street names.

2.3 Street names selection revision and preprocessing

The subsequent review of the resulting street names dataset uncovered several sources of error that needed to be addressed before continuing with the analysis. One issue was that some of the entries referred to places instead of people. This proved difficult to resolve at the beginning of the analysis, as surnames and place names can often be identical. As a result, disambiguation was generally only possible toward the end of the process, when no search could identify an individual corresponding to the name.

The second, more challenging issue was that, in many cases, the street name alone was insufficient to unambiguously identify the individual it referred to, particularly when it consisted of a common name or surname, or a combination of both.

Another consideration was that, in many instances, two different but very similar street names referred to the same person. For example, it is common to see both "Goya" and "Francisco de Goya" used to refer to the 18th-century Spanish painter. While this didn't present issues when searching in open databases, it did result in longer processing and quality-checking times. Identifying these corresponding instances would have helped reduce the number of strings to process.

In fact, the latter issue is a common challenge in address disambiguation. Many studies attempt to solve address matching using rule-based methods or text similarity, with more recent approaches relying on deep learning techniques to identify semantic similarities between different addresses, as seen in Xu et al. (2022). However, while this method has proven effective for matching common names, it is important to note that common and proper names do not operate within the same semantic framework.

Therefore, only minor changes were made to the dataset, such as removing punctuation, non-alphanumeric characters, and any unusual symbols.

2.4 Data search and extraction

Before initiating the data search, several databases were reviewed to determine which attributes could be extracted and to design the expected outcomes accordingly. The attributes selected for extraction were: name, full name, gender, description, occupation, date and place of birth, date of death (if any), occupation and URL. The extraction of these elements was carried out using different methods, depending on the data source.

2.4.1 *Linked Open Databases: Wikidata*

The search was conducted in two stages: first, we used the street name string to search for the corresponding entity codes via the “wbsearchentities” endpoint of the MediaWiki web service API. Once the candidate entity codes were identified, we utilized the Wikidata Query Service (WDQS) to retrieve the aforementioned attributes for each entity found in the previous stage, using SPARQL queries.

2.4.2 *Web Scraping in official sites: Electronic Biographic Dictionary of the Spanish Royal Society of History (DBH)*

The Spanish Royal Academy of History's search engine enables simple queries over its collections (covering over 50,000 individuals from the history of Spain and other countries) by embedding search terms or parameters directly into the URL. These searches yield a well-structured HTML page that is easy to scrape, from which the aforementioned attributes (excluding gender) can be extracted. Gender extraction is then processed by analyzing the person's name and the adjectives used to describe them in the description field.

Figure 1: screenshot of the heading of the results of a search in the DBH

2.4.3 *Web Scraping in Wikipedia*

As with the previous case, the street name was embedded as a search parameter directly into the URL. In some instances, we were able to extract the desired data from the information table that appears on the right side of each page. In other cases, the person's name only appeared in an annex or as secondary information, requiring a more detailed evaluation to locate the data we needed.

Proceedings of the International Cartographic Association, 7, 20, 2025 | https://doi.org/10.5194/ica-proc-7-20-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

2.5 Collected dataset evaluation and classification

In many cases, multiple candidate results were returned for a single person. For the searches conducted in Wikidata, we ranked the results based on the number of references to each candidate and selected the most popular one.

When this approach wasn't feasible, some results were discarded based on the candidate's description. For example, in one case the candidate was described as a terrorist; this kind of result were immediately disregarded due to the low likelihood of a street being named after a criminal.

In other cases, when there was no clear criterion to choose one result over the others, all the results were retained.

The final results show that most of the information was sourced from Wikidata, which served as the primary resource, while the other two sources were used only as backups when no results were available in Wikidata. The analysis also shows that, of all the streets included, only 20% were named after women, while approximately 80% were named after men. Additionally, it highlights that finding information about streets named after men is considerably easier, as over 50% of the cases involving streets named after women yielded no available information.

Source	Street names matched		
	Female	Male	Total
Not found	1153	2838	3991
WikiData	995	5122	6117
DBE	37	473	510
WikiPedia	55	411	466
Total general	2240	8844	11084

Table 1. Number of results per source and gender

It is worth noting that most of the streets for which information could not be found in the data sources are named after religious figures (such as saints, virgins, etc.).

2.6 Classification

The collected descriptions and occupations have been organized into seven distinct categories to identify the fields to which the streets are associated:

a) Culture and arts
b) Power, army and politics
c) Religion
d) Sport
e) Fiction and mythology
f) Science and investigation
g) Society: other kind of occupations not directly related to science or culture / arts (lawyers, entrepreneurs, professors, etc)

3. Results

The results reveal a clear imbalance in the gender of the individuals after whom the streets are named. In all municipalities of the region, male names significantly outnumber female ones. The few cases where there are more streets named after women than men are found in very small villages, where almost all the streets are named after religious figures.

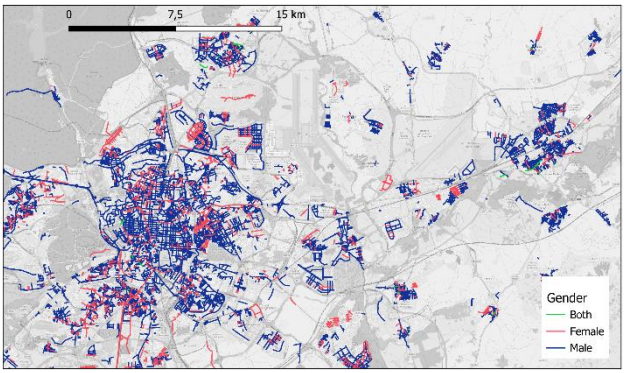

Figure 2: map showing distribution of streets named after males (blue) and females (red) in Madrid capital and the zone surrounding it to the East.

Regarding the distribution of street names by gender and their assigned category, the diagram below illustrates that the most common category for streets named after females is religion, followed by culture and power. For streets named after males, the most common category is culture, followed by religion and power.

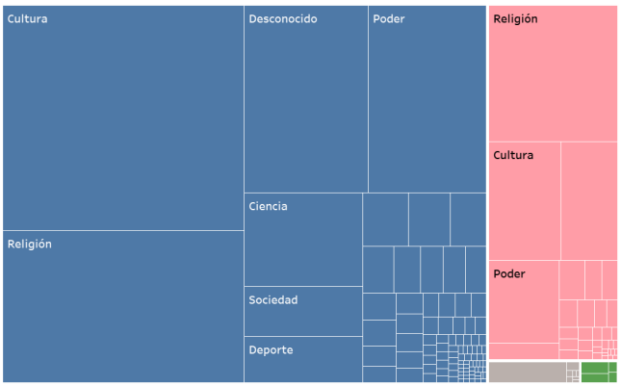

Figure 3. Area diagram showing the number of male (blue) and female (red) names in the Streets of the Region of Madrid. In the case of males, those linked to culture (followed by religion and power) prevail; in the case of women, those linked to religion outnumber the remaining areas (followed by culture and power).

However, an interesting difference emerges in the representation of power: while female figures are predominantly queens, members of royalty, and nobles from past centuries, male figures are more frequently associated with politics and military figures from the last decades.

The distribution of these three main categories in the city of Madrid is illustrated in Figures 4, 5, and 6 on the following page. It is also noteworthy to observe the spatial distribution of streets in the power and culture categories, where streets named after men tend to be located in more central areas, while streets named after women are often found on the periphery. This suggests that the latter are likely part of more recent urban developments.

Proceedings of the International Cartographic Association, 7, 20, 2025 | https://doi.org/10.5194/ica-proc-7-20-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

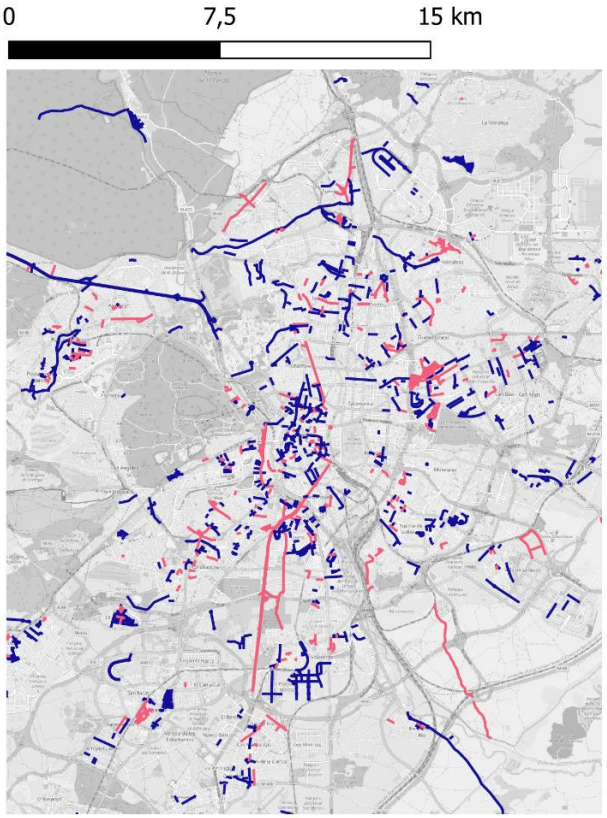

Figure 4: distribution of streets named after males (blue) and females (red) in central Madrid, classified in the "religion" category.

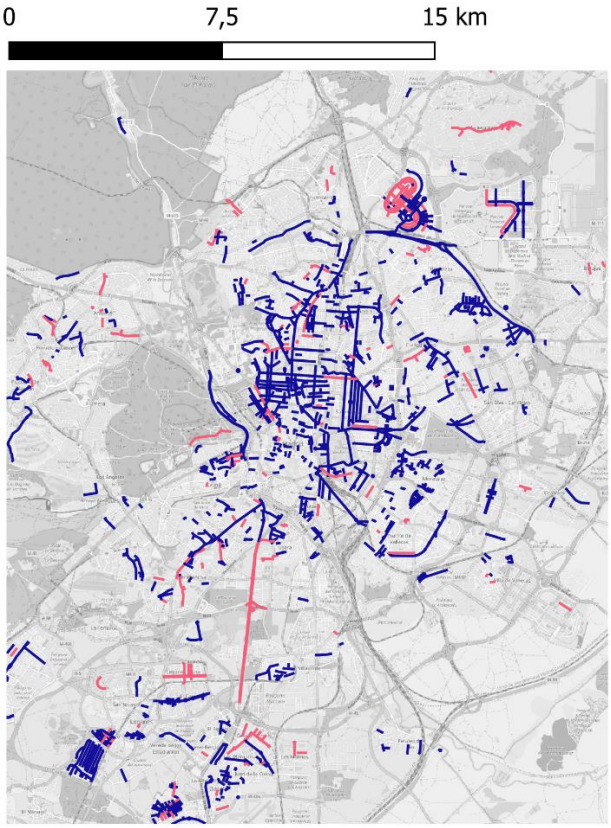

Figure 5: distribution of streets named after males (blue) and females (red) in central Madrid, classified in the "Power and Politics" category.

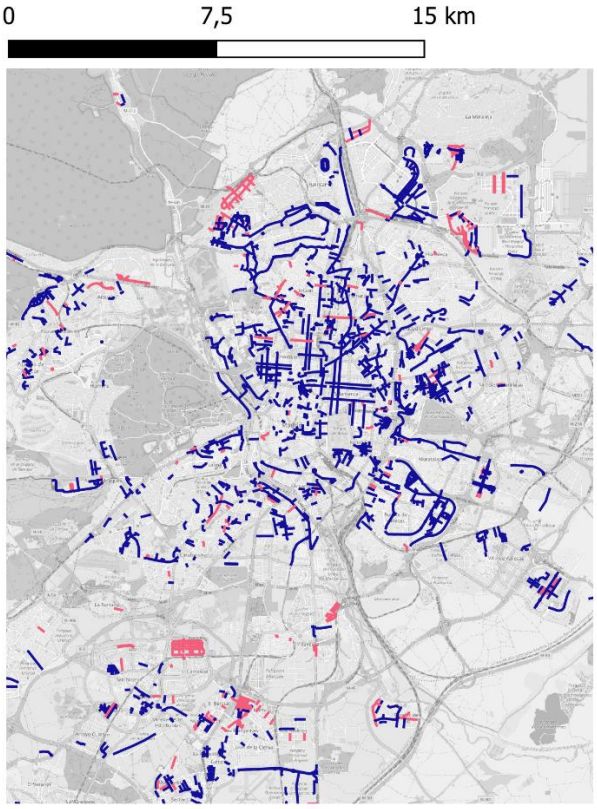

Figure 6: distribution of streets named after males (blue) and females (red) in central Madrid, classified in the "Culture" category.

4. Future work

On the short term, we are working to publish this project through an interactive Story Map that shows the workflow and intermediate results, and a dashboard that lets the user explore and interact with the data, helping create awareness about the patterns we follow as a society. Also, by publishing it we expect to foster map users' collaboration to detect eventual mistakes, to identify street names that should have been changed by law and to improve the knowledge we have about our history.

Building on the progress made so far, our goal is to further refine the named entity recognition process at the initial stage of the workflow and enhance the name-matching procedures to reduce the need for manual intervention. We aim to scale the methodology to cover the entire national territory, providing a comprehensive overview of the situation. Additionally, we plan to expand both the range and complexity of extractable attributes to address more sophisticated research questions.

On the other hand, we expect to improve the results in certain areas by exploring and leveraging existing Open Linked Databases in the field of Digital Humanities, Cultural Heritage and Art. Alexiev (2018) gives an in-depth analysis of the standards, ontologies and taxonomies developed in this fields.

Proceedings of the International Cartographic Association, 7, 20, 2025 | https://doi.org/10.5194/ica-proc-7-20-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

5. References

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S., 2021. Automated Linking of Historical Data. In: *Journal of Economic Literature* 59 (3): 865–918. DOI: 10.1257/jel.20201599

Alexiev, V., 2018. Museum Linked Open Data: Ontologies, Datasets, Projects. In: *Digital Presentation and Preservation of Cultural and Scientific Heritage*. 8. 19-50. https://doi.org/10.55630/dipp.2018.8.1.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Pio Carrino, C., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., & Villegas, M., 2022. MarIA: Spanish Language Models. In: *Procesamiento Del Lenguaje Natural*, *68*, 39-60. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405

Olmedo Ramos, J., 2007. El diccionario biográfico español de la Real Academia de la Historia. In: *Cercles: revista d'història cultural*, *núm. 10, p. 82-101*, https://raco.cat/index.php/Cercles/article/view/191234.

Miyakita, G., Leskinen, P., & Hyvönen, E. (2018). Using linked data for prosopographical research of historical persons: Case U.S. congress legislators. In A. Doulamis, E. Fink, M. Ioannides, R. Brumana, M. Wallace, P. Patias, & J. Martins (Eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection - 7th International Conference, EuroMed 2018, Proceedings* (pp. 150-162). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11197 LNCS). Springer Verlag. https://doi.org/10.1007/978-3-030-01765-1_18

Villar, J.J. (2022). Toponimy and urban history: creation, evolution and political manipulation of streets names in Bailén. In: *Locvber: Revista científica de patrimonio*, ISSN-e 2603-5847, vol. 6, 2022, pp. 23-77.

Woltjer, P. E., Zandhuis, I., Coret, B., Lindeman, M., Balkenende, J. D., Zijdeman, R. L., & Mourits, R. J., 2024. Persons in Context. A Model to Represent Observations and Reconstructions of Historical Persons in Linked Data. In: *Historical Life Course Studies*, 14, 105–125. https://doi.org/10.51964/hlcs19312

Xu, L., Mao, R., Zhang, C., Wang, Y., Zheng, X., Xue, X., & Xia, F., 2022. Deep Transfer Learning Model for Semantic Address Matching. In: *Applied Sciences*, *12*(19), 10110. https://doi.org/10.3390/app121910110