“Desambiguando” el género: cómo obtener género de nombres de manera sistemática, gratuita y consciente

Elvira González-Salmón

#YoSigoPublicando

28 noviembre 2023

¿A qué nos referimos con desambiguación* de género?

Lista de nombres

↓

"Algoritmo"

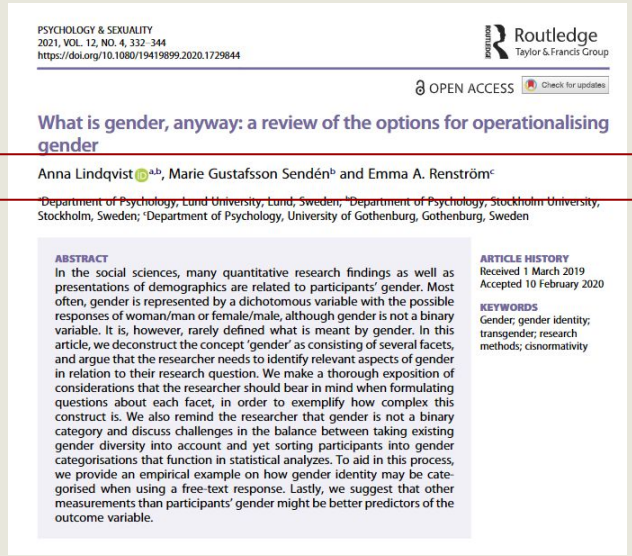
↓

Lista de nombres y género

↓

Estudios que tienen en cuenta el género

Ejemplo: bibliometría



PSYCHOLOGY & SEXUALITY
2021, VOL. 12, NO. 4, 332–344
https://doi.org/10.1080/19419899.2020.1729844

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

What is gender, anyway: a review of the options for operationalising gender

Anna Lindqvist[a,b], Marie Gustafsson Sendén[b] and Emma A. Renström[c]

[a]Department of Psychology, Lund University, Lund, Sweden; [b]Department of Psychology, Stockholm University, Stockholm, Sweden; [c]Department of Psychology, University of Gothenburg, Gothenburg, Sweden

ABSTRACT

In the social sciences, many quantitative research findings as well as presentations of demographics are related to participants' gender. Most often, gender is represented by a dichotomous variable with the possible responses of woman/man or female/male, although gender is not a binary variable. It is, however, rarely defined what is meant by gender. In this article, we deconstruct the concept 'gender' as consisting of several facets, and argue that the researcher needs to identify relevant aspects of gender in relation to their research question. We make a thorough exposition of considerations that the researcher should bear in mind when formulating questions about each facet, in order to exemplify how complex this construct is. We also remind the researcher that gender is not a binary category and discuss challenges in the balance between taking existing gender diversity into account and yet sorting participants into gender categorisations that function in statistical analyzes. To aid in this process, we provide an empirical example on how gender identity may be categorised when using a free-text response. Lastly, we suggest that other measurements than participants' gender might be better predictors of the outcome variable.

ARTICLE HISTORY
Received 1 March 2019
Accepted 10 February 2020

KEYWORDS
Gender; gender identity; transgender; research methods; cisnormativity

* O estimación, identificación, asignación, etc.

Organización del curso

1. Utilidad del curso
2. Cuestiones éticas
3. Funcionamiento básico
4. Cuestiones a considerar
5. Opciones
6. Casos prácticos
7. Bibliografía

Ej. NamSor

1. Utilidad del curso

1. Utilidad del curso: ¿Cuándo identificar el género de nombres de manera sistemática?

1. Número considerable de datos

2. No hay opción de autoidentificación
   - Encuesta

3. Es el único dato que tenemos (i.e. Face++, pronombres)

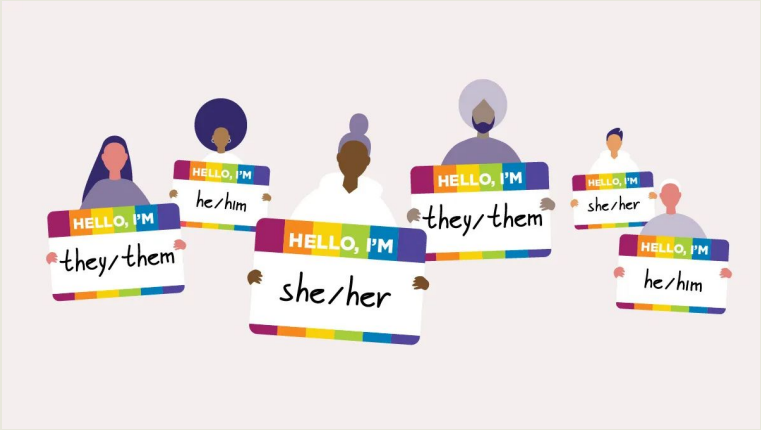
4. Cuando tiene sentido para nuestra investigación (Lindqvist et al., 2021)
   - ¿Queremos saber el género o aspectos físicos o por ejemplo género legal?
   - ¿Es realmente parte de la investigación?

Conclusión: evitarlo siempre que se pueda

2. Cuestiones éticas

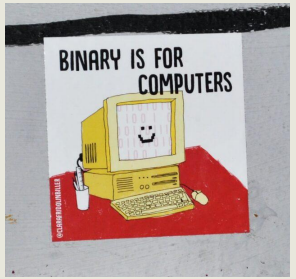
2. Cuestiones éticas

- Vemos la **probabilidad** de que un nombre sea asignado a un hombre o mujer, no asignamos género a personas individuales
- Habrá errores en esa probabilidad (*Misgendering*)
- Asume género de nombres - Refuerzo de estereotipos



2. Cuestiones éticas

- Imagen binaria del género (Frohard-Dourlent et al., 2016; Medeiros et al., 2020)
  - Understanding Nonbinary People: How to Be Respectful
  - ¿Qué es el género no binario?
- No sabemos cómo de grande es el error



> While the data collection for She Figures only considers sex-disaggregated data for men and women, it will be important to also consider non-binary gender for data collection in future publications, where possible. Non-binary is an umbrella term for gender identities that fall outside the gender binary of man or woman. This includes individuals whose gender identity is neither exclusively man nor woman, a combination of man and woman or between or beyond genders. The United Nations Economics Commission for Europe (UNECE)

She Figures, 2021

2. Cuestiones éticas

- Mucha gente trabajando en esto
- Importancia de los *disclaimers* y reconocimiento de limitaciones
  - Cada vez hay más
  - De “simply by looking at the full names of the researchers” a

Gender name-matching using Genderize.io. Our first gender classification step employed the Genderize.io API (https://genderize.io (Strømgren 2016); accessed 25 August 2019), which contains over 200,000 distinct first names and self-reported gender data from social media platforms across 79 countries and 89 languages. To determine the proportion of usage of any given-name for (fe)males, we sent three requests using either the country of last institutional affiliation as the locality, the US as locality, or no locality specified (in any of the records). These queries returned the number of occurrences of the given-name in Genderize.io's database based on locality and the proportion of occurrences for (fe)males (between 0 and 1). For given-names with diacritic characters (e.g., acute accent) for which no entries were identified, we re-sent the queries with the accent character replaced using the ustrnormalize function that normalizes composited Unicode code points to a decomposed form (e.g., Damián to Damian). Of the 20,989 unique given-names matched against the entire Genderize.io database, 56.22% appeared in the Genderize.io name corpus, with a median number of 15 entries with no locality specified (mean = 145.7), 10 with country of affiliation as locality (mean = 79.2), and 9 with the US as the locality (mean = 99.4). The gender assignment certainty (share of one gender among all occurrences) among these results is relatively high: for unique given-names with no locality specified, 93%, 89.1%, and 84.5%, respectively, have at least 70%, 85%, and 95% of Genderize.io records being one gender or the other, which (mostly) increases to 95.5%, 93.3%, and 90.4% (92.9%, 89.6%, and 85.4%) when country of affiliation (US) is used as locality. Nonetheless, even though using country of affiliation as the locality in the gender name matching may provide a more accurate classification given that the underlying data are drawn from national populations (e.g., national census or Facebook users), doing so might also limit the number of occurrences. For example, of the 41 countries with entries for the given-name Peter, 17 had zero records of Peter in the Genderized.io name list, in contrast

10

to 2,669 for the US and 738 for Great Britain. Hence, because gender assignment certainty could be less reliable when the number of entries is small, we checked the reliability of the Genderize.io gender classification using the following two criteria: gender assignment certainty higher than 0.7 and the number of entries for the given-name that is higher than the 10th percentile of other names in the same locality (more popular given-names) and a gender assignment certainty statistically different (at a 5% level) from 0.5 in a one-sample test of proportion given the number of Genderize.io entries. If either of these two criteria was satisfied for country of affiliation as locality (95.8%), then we used the gender assignment certainty. If both failed, we first checked the criteria when the US was used as locality (81.7%) and then when no locality was provided (28.7%). Based on this test, we excluded 3.29% of the unique given-names found in the Genderize.io corpus because of low confidence level.

Global Name Data project. The second step in our gender matching process drew on the gendered name list from the Global Name Data project (https://github.com/OpenGenderTracking/globalnamedata; accessed 3 September 2019), which comprises one list based on name and gender records from the US Social Security Administration and a second based on records from the UK Office of National Statistics, Northern Ireland Statistics and Research Administration, and Scotland General Register Office. These two name lists contain 91,321 and 34,347 unique given-names, respectively, together with binomial estimates of each observation's gender. Of all the unique given-names of top scientists, 58.65% (22,473) and 49.43% (18,940) were found in the US and UK lists, respectively. Nevertheless, for precision, we only considered binomial estimates higher than 0.7 (based on the proportion of names in either gender from all records) for which the 95% confidence interval did not include 0.5

11

(i.e., significance at a 5% level). Using this criterion, we excluded 756 and 2,186 unique given-names in each corresponding list but then found that only 0.3% of all the given-names fitting the significance criteria were classified in different genders across the two lists. We resolved this conflict by cross-checking the gender assignments for countries with a greater number of records with those of Genderize.io (which additionally considers surname location). Of the 21,126 given-names with a gender estimate from both databases, 3.48% (734) had a different gender classification (Genderize.io vs. Global Name Data: Pearson correlation = .935).

Other name lists. Next, in an approach similar to that of Larivière et al. (2013), we turned to online sources of country-specific gender name lists such as Wikipedia (https://en.wikipedia.org/wiki/Category:Feminine_given_names and https://en.wikipedia.org/wiki/Category:Masculine_given_names, accessed 8 September 2019) and Behind the Name (https://www.behindthename.com/names/gender/feminine and https://www.behindthename.com/names/gender/masculine, accessed 8 September 2019). Wikipedia contains 81 and 69 pages of masculine and feminine given-names based on country and language (e.g., French, Irish language) comprising 8,639 and 5,237 given-names, respectively. Behind the Name classifies names based on usage (e.g., Arabic, Biblical Latin, Greek mythology) for a total of 125,765 given-names (87,455 unique names) in 754 categories. For the remaining unclassified names, we first matched the names with the country-specific Wikipedia sources based on a scientist's country location and then double-checked against the Behind the Name list. To address three special cases, for scientists listing their country of affiliation as Russia, Czech Republic, Bulgaria, Latvia, Kazakhstan, Uzbekistan, Lithuania and Luxembourg with indeterminable gender based on given-name, we employed the suffixes rule (Larivière et al. 2013)

12

of assigning female (male) gender when the family name ends with -ova, -eva, or -ina (-ov, -ev, or –in). Likewise, for unclassified given-names with Japan as country of affiliation, we again followed (Larivière et al. 2013) by assigning female (male) gender to names ending with -a, -e, -i, -o, -ho, -ka, -ki, -ko, -mi, -na, -no, -ri, -sa, -ya, -yo, and -chi (-aki, -fumi, -go, -gu, -haru, -hei, -hide, -hiko, -hiro, -hisa, -ichi, -ji, -jiro, -ju, -kazu, -ku, -ma, -masa, -michi, -mitsu, -mu, -nari, -nobu, -nori, -rou, -ru, -shi, -shige, -su, -suke, -ta, -taka, -to, -tomo, -toshi, -yu, -zo, -zou, and –zu). Lastly, for the small number of observations with suffixes denoting generational titles (e.g., Sr., Jr., II, III, predominately scientists with US affiliation) with gender as yet undetermined, we assigned male gender across the board. The automatic gender classification outlined above, however, was somewhat inefficient for given-names affiliated with some East Asian countries; for example, China (only 32.9% assigned a gender), Taiwan (39.4%), Hong Kong (69.1%), Singapore (71.1%), and Korea (73.2%). Three research assistants thus manually classified the gender of scientists from these countries and that of other scientists not successfully categorized via earlier approaches. Manual coding of Chinese/Asian names was performed by two native speakers from China based on their own knowledge supplemented with Internet searches on the scientist's profile (e.g., departmental website) using full-name and institutional affiliation, while coding of other names was conducted by native English speakers from Australia. All the research assistants were instructed to determine scientist gender in ambiguous cases by searching Google Image and Facebook user accounts by either full or given-name.

Supplementary Results

Top-cited scientists based on citation records in 2017. The main findings in our paper were based on the top 100,000 most cited scientists during the period between 1996 and 2017; these are

13

3. Funcionamiento básico

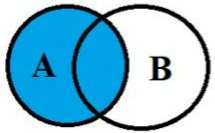
3. Funcionamiento básico

Juntar dos listas a partir de la columna que tienen en común (se añaden el resto de columnas)

Student ID	Name
1001	A
1002	B
1003	C
1004	D

+

Student ID	Department
1004	Mathematics
1005	Mathematics
1006	History
1007	Physics
1008	Computer Science



Student ID	Name	Department
1001	A	NULL
1002	B	NULL
1003	C	NULL
1004	D	Mathematics

3. Funcionamiento básico

Mi lista inicial

Nombre
Sara
Marta
Andrea
Federico
A.

Algoritmo

Nombre	Género
Sara	Female
Sonia	Mostly_female
Andrea	Mostly_female
Federico	Male
Pedro	Male
Andrea	Male

Mi lista final

Nombre	Género
Sara	Female
Marta	Unknown
Andrea	Mostly_female
Federico	Male
A.	Unknown

3. Funcionamiento básico

60% de identificados

El % aceptable de identificados para que tenga sentido los resultados dependen del caso

Muchos algoritmos dan % de probabilidad de que el género haya sido asignado correctamente, o tienen de posibilidades Male/Female/Mostly_male/Mostly_female (o similar)

Mi lista final		
Sara	Female	98%
Marta	Unknown	-
Andrea	Mostly_female	72%
Federico	Male	89%
A.	Unknown	-

4. Cuestiones a considerar

4. Cuestiones a considerar: Países

Mi lista inicial

Nombre	País
Sara	España
Marta	España
Andrea	Italia
Federico	Rusia
A.	Italia

Algoritmo

Nombre	País	Género
Sara	España	Female
Sonia	España	Female
Andrea	España	Mostly_female
Andrea	Italia	Male
Federico	España	Male
Federico	Rusia	Mostly_male
Pedro	España	Male
Alex	Estados Unidos	Unisex
Alex	España	Male

Mi lista final

Nombre	País	Género
Sara	España	Female
Marta	España	Unknown
Andrea	Italia	Male
Federico	Rusia	Mostly_male
A.	Italia	Unknown

4. Cuestiones a considerar: Países

- Nombres con género diferente en países diferentes (Ej. Andrea)
- Nombres eslavos, de Islandia, incluyen información de género en apellidos
- ¿Sobrerrepresentación de ciertos países?

Apellidos eslavos	
Male	**Female**
-ov -ev -in -ob -eb -yi -yj -ky -...	-ova -eva -ina -oba -eba -aya -aia -iha -...

4. Cuestiones a considerar: Origen de los datos

¿Sabemos de dónde vienen las listas de nombres que usamos?

- Nombres occidentales, problemas con Asia principalmente

- Sesgos de bases de datos (i.e., Wikipedia)

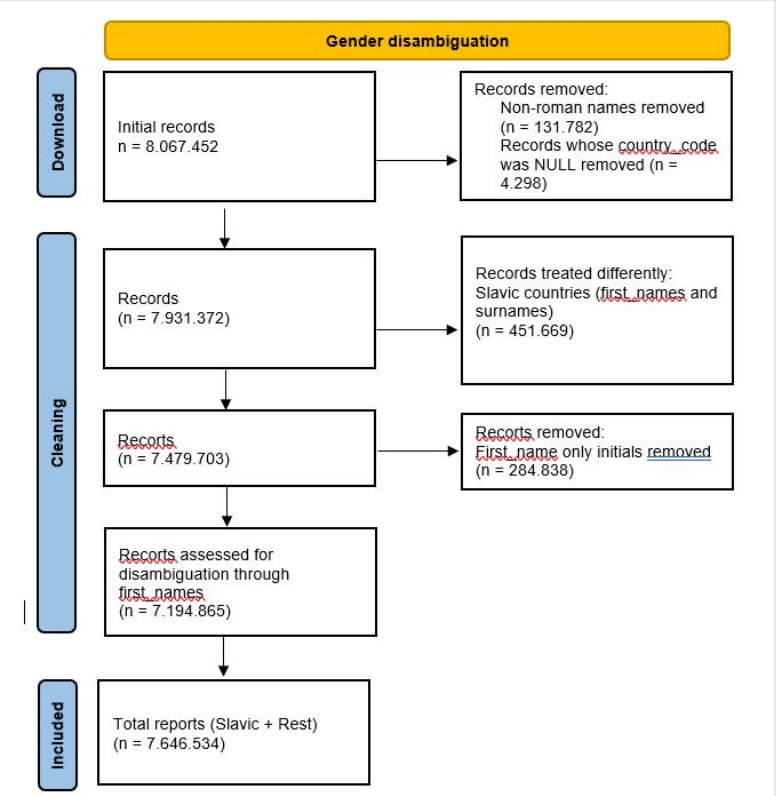
- ¿Datos (des)actualizados? (i.e. Nombre “Taylor” en EEUU)

4. Cuestiones a considerar: Datos limpios

Importante tener los datos limpios.

Problemas con:

- Iniciales
- Alfabeto no latino
- Segundos nombres
- Tildes y símbolos raros
- Considerar si queremos desambiguar todos los nombres por igual o por bloques



Esquema PRISMA

5. Opciones

¡No dejarse impresionar!
Muchos están repetidos

5. Opciones

	Gratuitos	Fuentes fiables	Plataforma	Actualizado	Nº nombres
Gender API	❌*	❌	Propia	Sí	6,084,389 de 190 países
Genderize.io	❌	⭐	Propia	Sí	4.079.646 de 188 países
Gender Guesser	❌	❌	Propia	Sí	9.000.000
Gender-guesser	✅	❌**	Python	No (2016)	44.568
NameAPI	❌*	❌	Propia	?	590.000 de 55 países
Gender-detector	✅	⭐	Python	No (2015)	125.000 de 2 países
SexMachine	✅	❌**	Python	No (2013)	44,568
World Gender Name Dictionary	✅	⭐⭐⭐	Python	Sí	25.000.000 de 195 territorios

*Free trial **Nam_dict.txt

5. Opciones

- Listas de nombres que nosotrxs *matcheamos* con Access, R, etc.
  - nam_dict.txt
  - Listas caseras
  - Listas desambiguadas de otra gente
- ChatGPT 3.5 ❌
- Usar más de un método (validación)

6. Casos prácticos

6. Casos prácticos: mis datos

	last_name	first_name	country_code
1	Woolf	Virginia	United Kingdom
2	Venturini	Aurora	Argentina
3	Sáez	Juanjo	Spain
4	Pereza	Elisa	Spain
5	Highsmith	Patricia	United States
6	López	Pedro López	Spain
7	Cavarero	Adriana	Italy
8	Yoshimoto	Banana	Japan
9	Innerarity	Daniel	Spain
10	Kulczycki	Emanuel	Poland
11	Harding	Sandra G.	United States
12	Ditlevsen	Tove	Denmark
13	Winsloe	Christa	Germany
14	Tokarczuk	Olga	Poland
15	Bicecci	Verónica Gerber	Mexico
16	Calderón	Javier	Spain
17	Khodyreva	Anastasia A.	Unknown
18	Vallejo	Irene	Spain
19	Hemingway	Ernest	United States
20	O'Farrell	Maggie	United Kingdom
21	Romero	Marga Sánchez	Spain

6. Caso práctico #1: Gender API

o

6. Caso práctico #2: World Gender Name Dictionary

¡Cuidado!

	first_name	country_code	level	gender	F	M
2	TT	SA	NaN	not found	not found	not found
6	I. A.	RU	3.0	M	0.0	1.0
0	F.	BR	NaN	not found	not found	not found
5	S.J.	US	NaN	not found	not found	not found
7	J	ES	3.0	M	0.0	1.0
1	S.M	US	NaN	not found	not found	not found
8	R. V.	RU	3.0	M	0.0	1.0
4	M.Z.	UA	NaN	not found	not found	not found
9	G.	FR	2.0	F	1.0	0.0
3	J.L.	GB	NaN	not found	not found	not found

Pasos:

1. Descargar
2. Abrir en Google Colab (o similar)
3. Añadir celda 1, 3, 10 y 11 de este (lo subiré a YSP)
4. Subir vuestro archivo (previamente limpiado) a la carpeta “example”

7. Bibliografía

7. Bibliografía: Cuestiones éticas

- Frohard-Dourlent, H., Dobson, S., Clark, B. A., Doull, M., & Saewyc, E. M. (2017). "I would have preferred more options": Accounting for non-binary youth in health research. *Nursing Inquiry, 24*(1), e12150. https://doi.org/10.1111/nin.12150
- González-Salmón, E & Robinson-Garcia, N. (2023). A call for transparency in gender assignment approaches. https://zenodo.org/records/10036331
- Heidari, S., Babor, T. F., De Castro, P., Tort, S., & Curno, M. (2016). Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. *Research integrity and peer review, 1*(1), 1-9.
- LaBrada, E. (2016). Categories We Die For: Ameliorating Gender in Analytic Feminist Philosophy. P*ublications of the Modern Language Association of America, 131*(2), 449–459. https://doi.org/10.1632/pmla.2016.131.2.449
- Lindqvist, A., Sendén, M. G., & Renström, E. A. (2021). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality, 12*(4), 332–344. https://doi.org/10.1080/19419899.2020.1729844
- Medeiros, M., Forest, B., & Öhberg, P. (2020). The Case for Non-Binary Gender Questions in Surveys. *PS: Political Science & Politics, 53*(1), 128–135. https://doi.org/10.1017/S1049096519001203
- Mihaljević, H., Tullney, M., Santamaría, L., & Steinfeldt, C. (2019). Reflections on Gender Analyses of Bibliographic Corpora. *Frontiers in Big Data*, *2*. https://www.frontiersin.org/articles/10.3389/fdata.2019.00029
- Rasmussen, K. C., Maier, E., Strauss, B. E., Durbin, M., Riesbeck, L., Wallach, A., Zamloot, V., & Erena, A. (2019). The Nonbinary Fraction: Looking Towards the Future of Gender Equity in Astronomy (arXiv:1907.04893; Version 1). arXiv. https://doi.org/10.48550/arXiv.1907.04893
- Stadler, G., Chesaniuk, M., Haering, S., Roseman, J., Straßburger, V. M., Martina, S., Aisha-Nusrat, A., Maisha, A., Kasia, B., Theda, B., Pichit, B., Marc, D., Sally, D. M., Ruth, D., Ilona, E., Marina, F., Paul, G., Denis, G., Ulrike, G., … Mine, W. (2023). Diversified innovations in the health sciences: Proposal for a Diversity Minimal Item Set (DiMIS). *Sustainable Chemistry and Pharmacy*, 33, 101072. https://doi.org/10.1016/j.scp.2023.101072

7. Bibliografía: Cuestiones técnicas

- Bérubé, N., Ghiasi, G., Sainte-Marie, M., & Larivière, V. (2020). Wiki-Gendersort: Automatic gender detection using first names in Wikipedia. *SocArXiv*. https://doi.org/10.31235/osf.io/ezw7p
- Blevins, C., & Mullen, L. (2015). Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction. *Digital Humanities Quarterly, 009*(3).
- Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 53–54. https://doi.org/10.1145/2872518.2889385
- Lax-Martinez, G., Raffo, J. D., & Saito, K. (2023). Identifying the gender of PCT inventors. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4434107
- Mryglod, O., Nazarovets, S., & Kozmenko, S. (2023). Peculiarities of gender disambiguation and ordering of non-English authors' names for Economic papers beyond core databases. *Journal of Data and Information Science, 8*(1), 72–89. https://doi.org/10.2478/jdis-2023-0001
- Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science, 4*, e156. https://doi.org/10.7717/peerj-cs.156
- Wais, K. (2016). Gender Prediction Methods Based on First Names with genderizeR. *The R Journal,* 8(1), 17. https://doi.org/10.32614/RJ-2016-002

7. Bibliografía: Ejemplos de asignación

- Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics, 125*(3), 2421–2447. Scopus. https://doi.org/10.1007/s11192-020-03733-w
- El-Ouahi, J., & Larivière, V. (2023). On the lack of women researchers in the Middle East and North Africa. *Scientometrics, 128*(8), 4321–4348. Scopus. https://doi.org/10.1007/s11192-023-04768-5
- Fell, C. B., & König, C. J. (2016). Is there a gender difference in scientific collaboration? A scientometric examination of co-authorships among industrial–organizational psychologists. *Scientometrics, 108*(1), 113–141. Scopus. https://doi.org/10.1007/s11192-016-1967-5
- Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature, 504*(7479), 211-213.
- Ma, Y., Teng, Y., Deng, Z., Liu, L., & Zhang, Y. (2023). Does writing style affect gender differences in the research performance of articles?: An empirical study of BERT-based textual sentiment analysis. *Scientometrics, 128*(4), 2105–2143. Scopus. https://doi.org/10.1007/s11192-023-04666-w

¡Muchas gracias! 🤠

¿Alguna pregunta?


elviragonzalez@go.ugr.es


@egonzalezsalmon