# Un dataset internacional acerca de nombres, género y frecuencias en DameGender

David Arroyo Menéndez, Madrid, Spain
{davidamgmail}.com

DAMEGENDER

## Abstract

La igualdad de género es el quinto objetivo de desarrollo sostenible (ODS) para Naciones Unidas[1].

Esta igualdad puede ser lograda midiendo, analizando datos y, creando buenas políticas con los resultados. Muchos estudios de género cuentan hombres y mujeres para explicar la posible desigualdad, por ejemplo, artículos de investigación, puestos de trabajo, calles, etc. El método tradicional de investigación es usar APIs comerciales con datos propietarios sin idea acerca de cómo los datos fueron recogidos. Los datos pueden también ser recogidos desde Wikipedia, estudios lingüísticos, sitios científicos, u oficinas estadísticas.

Este enfoque está basado en recoger Datasets Abiertos (Open Datasets) que incluyen nombre, género y frecuencia desde muchas instituciones estadísticas. Así, las tareas abordadas están basadas en unificar formatos, procesar datos y, crear pruebas para medir la precisión de los nuevos datasets.

El dataset usado cubre más de 20 países en el mundo occidental trayendo miles de nombres con una precisión de acierto mayor del 90%. Esto permitirá medir brecha de género a estudiantes y académicos interesados en el fenómeno sin costes y de una manera reproducible y más personas estarán contribuyendo a eliminar la brecha de género.

El Software Libre y los datos provistos por instituciones estadísticas hacen posible producir investigación reproducible por pares.

## 1 Introducción

Naciones Unidas tiene como objetivo reducir la brecha de género[2], para eso el primer paso es poder medir brecha de género ("if you cannot measure it, you cannot improve it" [Tho33]) para medir brecha de género resulta económico utilizar ordenadores ("Software Engineering Economics is an invaluable guide to determining software costs, applying the fundamental concepts of microeconomics to software engineering [B$^+$81]").

Además el software y los datos libres reducen más los costes, muchas personas e instituciones utilizan Software Libre como LibreOffice o Ubuntu GNU/Linux para evitar costes de licencias en productos similares como Microsoft Windows o Microsoft Word.

Esto creará una competición en el mercado con algún ganador, evitando pagos y generando beneficios desde una marca, tal y como ocurre en los navegadores con Firefox y/o Chrome.



[1] https://www.un.org/sustainabledevelopment/gender-equality/

[2] https://www.un.org/sustainabledevelopment/gender-equality/

A través del uso de los nombres personales, uno puede inferir el sexo o género de una persona en artículos académicos, libres, periódicos y muchas interacciones en Internet. Así, detectar género desde los nombres puede ser un camino estratégico para medir brecha de género, también.

Muchos usuarios y usuarias están, hoy en día, usando APIs tal y como Genderize, GenderAPI, Namsor, NameApi, Wikipedia, u otras soluciones libres (NLTK[LB02], R Gender, Gender Detector and Gender Computer[3]).

Las soluciones libres tradicionales tienen un bajo número de nombres debido al uso de un solo fichero que quizás solo pertenece a un país siendo software no mantenido a través del tiempo. Por otro lado, Wikipedia almacena pocos nombres por país.

Sin embargo, la brecha de género es un problema reconocido en Naciones Unidas y el mercado IT está liderando grandes desigualdades en economía y brecha de género. Este artículo presenta datos recogidos para asistir en la solución a un número de problemas (buscadores, inferiendo género en ficheros CSV, nombres en diferentes países, creación de nuevos datasets) encarados por la industria así como otros problemas que carecen de muchas soluciones industriales como contar hombres y mujeres en GitHub, listas de correo, etc.

Un estudio previo [KWL+16] mostraba una comparativa de datasets como un modo de mejorarla precisión, comparando herramientas que usan diferentes datasets públicos (SSA [4], IPUMS [5], namdict [6], etc).

Posibles aportaciones a este trabajo realizadas en este artículo serían aumentar el número de nombres y la atención a la diversidad en cuestiones como el género no binario o minorías culturales.

El dataset o quizás más bien DameGender (el software para crear el dataset) podría ser aplicado en artículos de las diversas disciplinas donde operan las herramientas de detección de género a partir del nombre [SGT+19]: ingeniería de software [VCS12], lingüística [Hut16, AZ09], ciencias sociales [HSFH18, MLA+11, NP17, dBA14]), ... Con vocación de ser puntuados muchos de ellos como de ciencia reproducible[Pen11] si se tienen las precacuciones correctas de la definición.

La estructura de este artículo es la siguiente: La sección 2 presenta la investigación principal midiendo brecha de género y herramientas de detección de género desde el nombre. La sección **??** da el vocabulario y la filosofía acerca de cómo elegir fuentes y encarar las problemáticas de diversidad construyendo un dataset La sección **??** explica una aplicación para este dataset: medir brecha de género en GNU/Linux La sección **??** es un resumen acerca de este enfoque y apunta a posibles trabajos futuros.

Las contribuciones de este artículo son

1. Una solución integrada en los diferentes campos de aplicación relativos a inferir género desde el nombre.
2. Una colección de datasets abiertos obtenidos desde fuentes estadísticas estandarizadas en un formato único.
3. Un nuevo estudio para contar hombres y mujeres en GNU/Linux.
4. Un enfoque basado en resultados reproducibles.

Muchos artículos relacionados muestran técnicas de aprendizaje automático aplicado, pero usando nuestro dataset libre mejoraría seguramente la precisión de los resultados contrastada con diversas fuentes abiertas como Wikipedia, Proyecto Gutenberg, Amazon, Forbes, instituciones deportivas, ...

También recogidas en DameGender y, mejorando el Estado del Arte en lo que respecta a Open Data en herramientas de detección de género a partir del nombre de manera significativa al cubrir más de 20 países (figura 1)

## 2 Estado del Arte

### 2.1 Acerca de la Brecha de Género

Reducir la brecha de género se refiere a igualdad entre hombres y mujeres y políticas de no discriminación. El género se refiere al sexo de la persona determinado en su nacimiento, aunque éste pueda cambiarse en algún momento de su vida. Las discusiones acerca de las definiciones de género se refieren a estos problemas. Sin embargo, hay un consenso determinando de género, frecuencia y nombres con estadísticas oficiales entregadas por las instituciones de los estados.

Medir brecha de género requiere un conjunto de indicadores. Global Gender Gap Report [CPSZ22] han sido propuesto en economía, salud, educación y política. El sitio de Naciones Unidas[7] está mostrando indicadores

[3] https://github.com/tue-mdse/genderComputer
[4] https://www.ssa.gov/oact/babynames/limits.html
[5] https://usa.ipums.org/usa-action/variables/NAMEFRST
[6] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt
[7] https://www.unwomen.org/

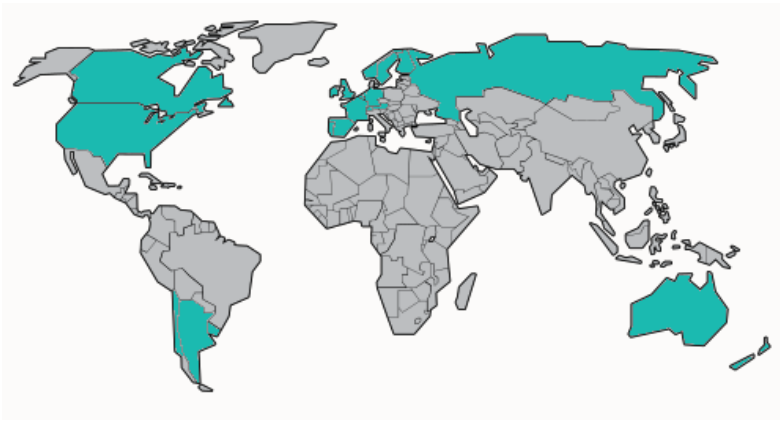

Figure 1: Green (Countries with Open Data provided by Statistical Institutions), White (Countries without Open Data or don't reproduced by the author). Thanks to wikimedia by the mapamundi base source `https://commons.wikimedia.org/wiki/File:Simple_world_map_edit.svg`

para medir disparidades como leyes, educación, mortalidad maternal, participación política, pobreza, trabajo doméstico, paridad de género en el trabajo, acceso a la economía, acceso a estudios y/o trabajo, violencia contra las mujeres, justicia climática, acceso a justicia, salud, etc.

Es posible crear decisiones de impacto en una cuestión a través de resultados de investigación que toman estos indicadores en consideración. Por ejemplo, Miyake y otros autores [MKSF+10] concluyeron que creando afirmaciones acerca de valores éticos la brecha de género se reducía en los colegios.

Relativo a medir brecha de género en investigación social, Bimber [Bim00] presentaba dos factores que afectan a la brecha de género en Internet (acceso y uso), razones de socioeconomía y género en una encuesta que recoge datos a través de varios años.

### 2.2 Contando hombres y mujeres en Internet. ¿Por qué? ¿Dónde?

El software DameGender focaliza en recuperar datos desde fuentes de datos diversas acerca de género y nombres. Lugares como GitHub, Wikipedia, APIs, periódicos, websites en general, listas de correo, ... Veamos cómo se estaban haciendo estos trabajos.

Por ejemplo, un científico social estudiando brecha de género en periodismo puede contar hombres y mujeres en Twitter. Mientras tanto, algún artículo de ciencias de la computación puede estar mejorando la detección de género en imágenes de caras, en nicknames, o simplemente en nombres.

Burger y otros [BHKZ11] presentaron varias configuraciones de un clasificador independiente de lenguaje para predecir el género de usuarios Twitter. El dataset usado para la construcción y evaluación de estos clasificadores se crea con usuarios Twitter que llegaban a tener blog en su perfil.

Este tipo de estudios permite analizar demografía de población Twitter [MLA+11], siendo del sexo una variable clave en cualquier estudio de demografía. En este caso, el dataset escogido fue el del censo de Estados Unidos que funciona razonablemente bien para nombres occidentales.

Wagner y otros [WGJS15] analizaban la brecha de género en Wikipedia mostrando evidencia de pequeñas formas de desigualdad explicando cómo resolver estas evidencias. Para medir la desigualdad de género se han desarrollado las siguientes bias: visibilidad, léxico (ej: palabras discriminatorias para mujeres), estructural y de cobertura.

Un buen número de personas billonarias puntúan en Forbes como de Ciencias de la Computación y los repositorios públicos nos permiten entender tendencias de inclusión o no de mujeres, así como listados públicos de puestos de trabajo, directivos, ... Las herramientas de detección de género a partir del nombre. Arjona y otros [RRGBD16] publicaron en 2016 una encuesta de 2000 contribuidores en el que la participación feminina iría del 2% al 5%. Recientemente, Zacchiroli un estudio longitudinal acerca de hombres y mujeres, que revelaba que las mujeres podrían ser el 8% en contribuciones de commits de código en repositorios públicos. Zacchiroli [Zac20] utilizó genderguesser para inferir género desde el nombre que a su vez usa namdict [8] como dataset. Vasilescu y

[8] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt

otros [VPR+15] exploraron StackOverflow un sitio muy popular de preguntas y respuestas concluyendo que los hombres representan la vasta mayoría de contribuciones.

Relativo a la brecha de género en ciencia Cassidy R. Sugimoto y otros [LNG+13] presenta un análisis bibliométrico confirmando que los desequilibrios de género persisten en investigación a través del mundo entero. Holman y otros [HSFH18] presentaban un código en R usando la API de genderize y proporciona un buen enfoque acerca de cómo calcular brecha de género.

### 2.3 Enfoques Automáticos para Inferir Género

Hay varios caminos para inferir género desde fuentes de Internet: texto escrito a mano, imágenes, documentos y nombres.

Liwicki y otros [LSB11] presentaba un método de inferir género desde textos escritos a mano con un 67.5%.

Gallagher y otros [GC08] presentaba un trabajo que detecta género a partir de una combinación de imágenes, edad e información cultural provista por primeros nombres con ejemplos no etiquetados y con una precisión del 60%.

Argamon y otros [AKFS03] explica que las mujeres usan muchos más pronombres que los hombres, mientras que los hombres usan más modificadores de nombres (determinantes, o adjetivos). Koppel y otros [KAS02] presentaban un sistema de clasificación de documentos con una precisión del 80 % aproximadamente. Cheng y otros [CCS11] muestran una selección de características (features) y un modelo construido usando aprendizaje automático generando un acierto del 85.1 % para identificar género desde el texto.

### 2.4 Infiriendo género desde nombres

Las herramientas usadas para inferir género desde un nombre están normalmente basados en datasets que, como mínimo, incluyen género.

Liu y otros [LR13] presentaba un método para inferir género desde los primeros nombres en Twitter, el dataset estaba codificado a mano por acuerdo entre tres trabajadores de Amazon con 50.000 usuarios Twitter seleccionados aleatoriamente con solo 12681 etiquetas de género. El objetivo de este estudio era determinar el valor incremental de usar el nombre de usuario (username) como una característica (feature) en la inferencia de género basado en tweets.

Mueller y otros [MS16] presentaban un trabajo de cómo inferir género en Twitter. Ellos usaban el dataset namdict y el censo de los Estados Unidos como datasets. Las características eran "número de consonantes", "número de vocales", "número de sílabas", "número de vocales bouba", "número de consonantes kiki", "número de vocales kiki". El modelo de clasificación era creado usando SVM.

### 2.5 Ideas Relacionadas

Ambekar y otros [AWM+09] presentaban un sistema de software libre para clasificar nombre y etnicidad usando aprendizaje automático para extraer una lista de nombres desde Wikipedia. Un trabajo más reciente guiado por Rodríguez Pérez y otros [NRN21], en el que fue presentado NamPrism dando ideas frescas acercar de clasificar razas y, siendo aplicado a repositorios de software masivos.

Bollegala y otros [BMI10] presentaba otro enfoque que usaba basado en patrones léxicos para extraer aliases de un nombre dado, con un conjunto de nombres y sus aliases como datos de entrenamiento. Los aliases candidatos entran en un ranking con varias puntuaciones. Support Vector Machine (SVM) fue usado para construir la función ranking.

### 2.6 Estándares Relacionados

ISO/IEC 5218 propone la siguiente norma acerca de la codificación de género: "0 como conocido", "1 como masculino", "2 como femenino" y "9 como no aplicable".

El RFC 6350 (vCard) tiene estas categorías "m como masculino", "f como femenino", "0 como otro", "n como no aplicable" y "u como indefinido (undefined)". Basado en este estándar, con respecto a publicación web se puede usar el microformato h_card en el contexto de escribir interfaces de formularios web considerando lecturas de w3.

### 2.7 Resumen

El primer nombre de un asunto es el factor clave usado para determinar género en el Estado del Arte de las herramientas de inferencia de género. Sin embargo, en muchos contextos hay más características: apellidos, texto, imágenes, nicknames, etc.

El primer nombre puede ser útil para inferir otras cuestiones tal y como raza, etnicidad, o cultura, también.

El aprendizaje automático y la selección de características previas está siendo usado en muchos trabajos, aunque hay una discusión tal que es el mejor enfoque.

## 3 Diseño

### 3.1 Verdad y Falsedad en Nombres, Género y Frecuencia

Usar nombre, género y frecuencia es una idea comúnmente admitida en vez de otras enfoques como nombre y género, o nombre y varios grados de masculinidad y feminidad. Esto es así, porque hay bastante industria e instituciones que utilizan esta manera de hacer las cosas. Así, las frecuencias nos permiten calcular porcentajes de masculinidad y feminidad en cada nombre.

Con respecto, a veracidad en las datos DameGender permite elegir a sus usuarios confiar en empresas o instituciones para inferir el género de cualquier nombre.

Y, se aportan datasets públicos construidos desde instituciones estadísticas que lo permiten. Este trabajo de observación de datasets de instituciones oficiales tiene el potencial de favorecer la participación ciudadana ante cambios desde los gobiernos.

### 3.2 Género, Lenguaje, Nación y Diversidad

Hay reglas y excepciones en diferentes lenguajes para predecir si un nombre es masculino o femenino. Por ejemplo, en español o inglés, hay más nombres que finalizan con “a” clasificadas como femeninos que, como masculinos. Sin embargo, Andrea es femenino en España y, masculino en Italia. Así, es útil comprender el lenguaje y cultura asociado con un nombre. El lenguaje nos acerca a naciones, pero con diferencias, por ejemplo, en España hay varios idiomas: vasco, catalán, castellano (español) y gallego. Español es la lengua oficial más utilizada en España, pero también en otros países como Argentina, Méjico, Ecuador, Bolivia, ...

En DameGender se han juntado los datasets oficiales (los entregados por instituciones estadísticas) para generar datasets de lenguajes. En lenguajes donde no había instituciones estadísticas hemos valorado utilizar nombres descargados de Wikipedia descubriendo una baja cantidad de nombres una no muy buena calidad.

### 3.3 Obteniendo datos y datasets para DameGender

DameGender unificó los diferentes formatos descargados desde oficinas estadísticas oficiales para nombre, género y frecuencia en los siguientes países: Argentina, Austria, Australia, Bélgica, Canadá, Suiza, Alemania, Dinamarca, España, Finlandia, Francia, Gran Bretaña, Irlanda, Islandia, Noruega, Nueva Zelanda, Méjico, Portugal, Rusia, Eslovenia, Suecia, Estados Unidos de América y Uruguay.

Se han desarrollado diferentes scripts para la descarga de datasets:

- donwloadcsv.py: para la descarga de ficheros CSV desde las APIs
- donwloadjson.py: para la descarga de ficheros JSON desde las APIs
- get-wikidata-names.py: para la descarga de nombres desde Wikipedia
- get-wikidata-surnames.py: para la descarga de apellidos desde Wikipedia
- orig2.py: para la descarga de datasets desde instituciones estadísticas oficiales

Una posible mejora a este diseño sería unificar todas estas funcionalidades en un solo comando, por ejemplo, orig2.py

## Acknowledgments

We would like to thank: the statistical institutions by release of the open datasets about names, gender and frequency. Luz Galvis for the software contributions, Daniel Izquierdo and Laura Arjona for starting this research field at URJC all those working with Jesús González Barahona and Gregorio Robles.

## References


[AKFS03] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346, 2003.

[AWM+09] Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 49–58, 2009.

[AZ09] Abdul Wahed Qasem Ghaleb Al-Zumor. A socio-cultural and linguistic analysis of yemeni arabic personal names. *GEMA: Online Journal of Language Studies*, 9(2):15–27, 2009.

[B+81] Boehm Barry et al. Software engineering economics. *New York*, 197, 1981.

[BHKZ11] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[Bim00] Bruce Bimber. Measuring the gender gap on the internet. *Social science quarterly*, pages 868–876, 2000.

[BMI10] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Automatic discovery of personal name aliases from the web. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):831–844, 2010.

[CCS11] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[CPSZ22] Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World inequality report 2022*. Harvard University Press, 2022.

[dBA14] Clara Sainz de Baranda Andújar. El género de los protagonistas en la información deportiva (1979-2010): noticias y titulares/the gender of the main characters in sports reporting (1979-2010): News and headlines. *Estudios sobre el mensaje periodístico*, 20(2):1225, 2014.

[GC08] Andrew C Gallagher and Tsuhan Chen. Estimating age, gender, and identity using first name priors. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[HSFH18] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. The gender gap in science: How long until women are equally represented? *PLoS biology*, 16(4):e2004956, 2018.

[Hut16] ¿ Matthew Hutson. The gender of names. *Scientific American Mind*, 27(4):14–14, 2016.

[KAS02] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[KWL+16] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54, 2016.

[LB02] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[LNG+13] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.

[LR13] Wendy Liu and Derek Ruths. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*, 2013.

[LSB11] Marcus Liwicki, Andreas Schlapbach, and Horst Bunke. Automatic gender detection using on-line and off-line information. *Pattern Analysis and Applications*, 14(1):87–92, 2011.

[MKSF+10] Akira Miyake, Lauren E Kost-Smith, Noah D Finkelstein, Steven J Pollock, Geoffrey L Cohen, and Tiffany A Ito. Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008):1234–1237, 2010.

[MLA+11] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[MS16] Juergen Mueller and Gerd Stumme. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, pages 1–8, 2016.

[NP17] Mari K Niemi and Ville Pitkänen. Gendered use of experts in the media: Analysis of the gender gap in finnish news journalism. *Public Understanding of Science*, 26(3):355–368, 2017.

[NRN21] Reza Nadri, Gema Rodriguezperez, and Meiyappan Nagappan. On the relationship between the developer's perceptible race and ethnicity and the evaluation of contributions in oss. *IEEE Transactions on Software Engineering*, 2021.

[Pen11] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[RRGBD16] Gregorio Robles, Laura Arjona Reina, Jesús M. González-Barahona, and Santiago Dueñas Domínguez. Women in free/libre/open source software: The situation in the 2010s. In Kevin Crowston, Imed Hammouda, Björn Lundell, Gregorio Robles, Jonas Gamalielsson, and Juho Lindman, editors, *Open Source Systems: Integrating Communities*, pages 163–173, Cham, 2016. Springer International Publishing.

[SGT+19] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[Tho33] W Thompson. Electrical units of measurement, popular lectures, 1833.

[VCS12] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*, pages 332–338. IEEE, 2012.

[VPR+15] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in github teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3789–3798. ACM, 2015.

[WGJS15] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

[Zac20] Stefano Zacchiroli. Gender differences in public code contributions: a 50-year perspective. *IEEE Software*, 38(2):45–50, 2020.