DivinWD: Exploring the Diversity of Scientific Publications in Wikidata

Zeno Saletti[1], **Cristian Consonni**[2,*], **Pedro Frau-Amar**[3,*], **Emilia Gómez**[3,*]

[1]University of Trento, DISI, Trento, Italy
[2]European Commission, Joint Research Centre (JRC), Ispra, Italy
[3]European Commission, Joint Research Centre (JRC), Seville, Spain
zeno.saletti@studenti.unitn.it, cristian.consonni@acm.org, pedro.frau-amar@ext.ec.europa.eu,
emilia.gomez-gutierrez@ec.europa.eu

Abstract

Wikidata has emerged as a major open repository of scholarly metadata, yet its characteristics and limitations as a source for diversity analyses are not well documented. We present *DivinWD* (*DIVersity IN WikiData*), a curated dataset consisting of more than 23 million triples, produced via a fully open-source processing pipeline to enable the study of diversity in scientific publications represented in Wikidata. The dataset comprises over 1.2 million scholarly articles published between 2010 and 2024, enriched by integrating Wikidata with 5 external bibliographic sources–Crossref, Dimensions, OpenAlex, Scopus, and Semantic Scholar–and augmenting them using the Genderize API, to enrich metadata on language, field of study, authorship, gender, geographic origin, and institutional affiliation. Our analysis documents systematic coverage biases and infrastructural artifacts affecting Wikidata's scholarly content, highlighting important considerations for reuse. By releasing the dataset and pipeline, this work provides a transparent foundation for future research on diversity in science and for the development and evaluation of open, reproducible bibliometric indicators.

Code — https://github.com/DivinWD/dataset-resources
Datasets — https://zenodo.org/records/18234750

1 Introduction

The advent of the Semantic Web has transformed how bibliographic metadata informs our understanding of scientific output, yet much of this infrastructure remains fragmented across siloed systems or locked behind proprietary barriers. These limitations are especially pressing as artificial intelligence (AI) applications increasingly mediate access to knowledge: the quality and openness of underlying data sources are critical to building equitable AI and trustworthy information ecosystems.

Diversity and representation in science are prerequisites for such systems. The High-Level Expert Group on AI of the European Commission has identified diversity, fairness, and non-discrimination as central requirements for trustworthy AI (HLEGAI 2019), echoing fundamental rights enshrined in the Universal Declaration of Human Rights (Article 7, United Nations (1948)) and the Charter of Fundamental Rights of the European Union (Title III, European Union (2012)). These concerns are underscored by high-profile journalistic investigations revealing discriminatory outcomes in machine learning (Small 2023; Angwin et al. 2016). Research has also shows that diverse teams outperform homogeneous ones (Swartz et al. 2019; AlShebli, Rahwan, and Woon 2018; Freeman and Huang 2014) and produce more novel, highly cited research (Yang et al. 2022; Adams 2013), yet persistent inequalities continue to shape scientific participation (Oswald and Ostojic 2020). Monitoring the scientific record can therefore serve as a tool that can be used to foster inclusivity and to evaluate the effectiveness of diversity, equity and inclusion (DEI) initiatives.

Despite the increasing availability of bibliographic metadata, existing infrastructures remain fragmented across heterogeneous sources with varying access conditions and limited interoperability. As a result, monitoring diversity in science often requires substantial data integration efforts, limiting its scalability and reproducibility. Against this backdrop, Wikidata[1]–a free/libre collaborative knowledge graph, launched in 2012 as a platform for structured data in the Wikimedia ecosystem–has become a cornerstone of the Linked Open Data (LOD) ecosystem. Wikidata is freely accessible, multilingual, it is continuously curated by volunteers and it has grown into a vast knowledge graph; as of December 2025, it includes over 117 million items, among them more than 45 million scientific publications. Initiatives such as *WikiCite*,[2] together with tools like Scholia[3] (Nielsen, Mietchen, and Willighagen 2017; Willighagen et al. 2026) and large-scale data imports, have positioned Wikidata as a promising foundation for open bibliographic repositories.

Wikidata is embedded in an open, multilingual ecosystem and provides structured links between entities such as authors, institutions, and demographic attributes. These characteristics make it a promising foundation for building flex-

[*]*Disclaimer*: The view expressed in this paper is purely that of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

[1]https://www.wikidata.org/, accessed 2026-04-10.
[2]http://wikicite.org/, accessed 2026-04-10.
[3]https://scholia.toolforge.org/, accessed 2026-04-10.

ible and reproducible pipelines that can be adopted by scientific communities of varying sizes and resources. However, its collaborative nature and heterogeneous data ingestion processes raise important questions about completeness, consistency, and bias. This gap between potential and current understanding motivates our study.

We investigate these opportunities through *DivinWD* (*DIVersity IN WikiData*). In this paper, we release a curated dataset of scholarly articles and authors from Wikidata, together with the fully-reproducible pipeline for its computation. With *DivinWD* we answer the following questions: yo what extent can Wikidata records be successfully mapped to other established bibliographic infrastructures? What kinds of biases emerge when Wikidata is used as a lens for studying the diversity of scientific publications?

The main contributions of this paper are as follows:

1. we produce a curated bibliographic dataset, *DivinWD* , collating data from Wikidata and five major bibliographic sources–Crossref, Dimensions, OpenAlex, Scopus, and Semantic Scholar– and we augment the data inferring the field of study of each work and the gender and nationality of authors when missing;
2. we document a reproducible pipeline for processing and analyzing bibliographic data in Wikidata and we release it fully open-source;
3. as a use case, we explore several diversity dimensions for articles (language, field of study) and authors (gender, geographic origin, and affiliation) and we compute established indicators of diversity;
4. we provide insights for future inclusivity initiatives in scientific publishing.

The remainder of the paper is organized as follows: in Section 2 we introduce Wikidata and the other data sources we have used in this study; Section 4 describes our data processing pipeline. We presents our dataset in Section 6 and we compute several diversity attributes from this dataset. Finally, we discuss some lessons learned for future bibliometric initiatives regarding diversity in the Section 7 and we conclude the paper in Section 8.

2 Background

Wikidata. As described by Erxleben et al. (2014), Wikidata consists of *entities*, defined as subjects outlined by structured data. The data model of Wikidata separates entities into a few macro-categories; the most important ones are *items*, which are intuitively objects and classes, and *properties*, which describe entities by connecting them. Entities are associated with a unique identifier: *QIDs* for items, composed of the letter "Q" followed by digits, and *PIDs* for properties, composed of the letter "P" followed by digits. Human-readable data, which comprise labels, aliases, and descriptions, are referred to as *terms*; entities can have terms in multiple languages. The core of an entity is represented by a list of *statements*, each consisting of a property and a value, which may be an entity or another data type, including special values to denote "none" (no value) and "some" (the value is unknown). Depending on the definition of the property, Wikidata may issue warnings about malformed statements; however, constraints on property values are not strictly enforced, allowing for maximum flexibility. This is useful for improving property definitions against modeling challenges, but it is also prone to actual mistakes and inconsistencies. Furthermore, properties alone are not always sufficient to model complex scenarios, and users can add contextual information to statements in the form of *qualifiers*, which are property-value pairs that further describe the statement, and *references* that support the claim by linking it to one or more external sources. Some statements have properties designed to take external identifiers as values. These statements are collectively known as *identifiers*, and they usually link Wikidata items to corresponding objects in other databases, such as the Library of Congress authority ID or IMDb ID. Finally, statements are categorized according to a system of three ranks: normal, preferred, and deprecated. Ranks serve many purposes, e.g., to mark the most updated statements or those that best represent community consensus; conversely, deprecated statements can be kept for historical purposes. In this work, we consider only the best-ranked statements, leaving out deprecated ones.

Additional Bibliographic Data Sources. To complement Wikidata, we integrate multiple bibliographic data sources, including both open (Crossref, OpenAlex) and proprietary (Semantic Scholar, Dimensions, Scopus) database. These sources were selected based on their availability for large-scale data access and analysis. Crossref and OpenAlex allow us to fill missing attributes such as language and abstract information. Dimensions and Scopus support the validation of author lists and consistency checks. Semantic Scholar contributes field-of-study classifications, enabling the derivation of disciplinary information when absent in Wikidata. Finally, the Research Organization Registry (ROR) provides standardized typologies for institutions, allowing us to harmonize affiliation data. Together, these sources enable the enrichment of Wikidata records and support our analysis. Further details on these data sources and their governance are provided in Appendix A.

3 Related Work

DivinAI. This analysis builds on the DivinAI project (Gómez et al. 2024; Hupont et al. 2022; Freire, Porcaro, and Gómez 2021), which aimed to establish diversity metrics for monitoring gender, nationality, and institutional affiliations of individuals involved in artificial intelligence (AI) research and conferences, with a particular focus on the fields of recommender systems (Porcaro et al. 2023) and affective computing (Hupont et al. 2024, 2021). While our approach is aligned with DivinAI's focus on these dimensions, we differ in data sources: all metrics are derived from Wikidata rather than manually labeled data obtained from the articles.

Diversity in Science. The topic of diversity in science has become more prominent in recent years, but the amount of attention that has been dedicated to this topic varies across fields. Furthermore, diversity can refer to different aspects of the research process, from content (e.g. the diversity of topics or problems studied), to the characteristics of scientists themselves, of scientific collaborations, or diversity initiatives in conferences and academic communities. Previous work has found that research collaboration are often homofilic (AlShebli, Rahwan, and Woon 2018; Freeman and Huang 2015). Studies in geoscience reveal persistent gender disparities, with women under-represented in leadership roles such as editorial boards and conference organization (Vila-Concejo et al. 2018), and black women even more under-represented than white women (King et al. 2018; Dutt 2020); similar patterns are reported in STEM (Shishkova et al. 2017) and medicine (Spector et al. 2019; Partiali et al. 2021; Dijksterhuis et al. 2020). For research in the affective computing field, refer to (Hupont et al. 2024). A summary of diversity-related initiatives is provided in Table 1 of Gómez et al. (2024).

Bias and Diversity in Open, Collaborative Projects. The topic of diversity in collaborative projects has received attention is several scientific communities over the last few years. For example, geographic (Rossi and Zacchiroli 2022) or gender (Trinkenreich et al. 2022) diversity of free/open source contributors. Wikipedia has been shown to suffer a gender disparity in the number of contributors (Antin et al. 2011), which also translate to cultural and gender-based disparities in content coverage, quality, and prominence (Wagner et al. 2015, 2016; Pfeil, Zaphiris, and Ang 2006). A prominent area of investigation is the under-representation of women in Wikipedia biographies, with studies demonstrating a persistent male dominance (Yunus, Chen, and Demartini 2025; Patel et al. 2024).

Reliability, Bias and Diversity in Wikidata. Mora-Cantallops et al. have systematically analyzed the research literature about Wikidata and found that, while the study of the platform is still in the early stages, it is dominated by European researchers (Mora-Cantallops et al. 2019). Piscopo and Simperl (2019) surveyed 28 papers about data quality in Wikidata and found that completeness, consistency, and accuracy of data were the most studied dimensions. Similarly to Wikipedia, patterns of under-representation of women have also been found in Wikidata (Conroy 2023; Zhang and Terveen 2021; Konieczny and Klein 2018; Klein et al. 2016). Recent works by Paolini, Metilli and collaborators as part of the *WiGeDi* (Wikidata Gender Diversity) project (Metilli and Paolini 2023) have explored how the Wikidata community has approached the issue of modeling gender data, noting that over time the interpretation of gender has become more inclusive (Melis et al. 2025) and has been influenced by external events (Melis et al. 2024).

4 Methods

Responsible Data Collection Practices

Collecting data for studies of this kind raises important ethical considerations and risks. Many of the issues raised by Hupont et al. (2021) also apply here. Specifically, studies computing diversity metrics must balance the need to monitor diversity in science with strong privacy protection and careful consideration of how such data might be misused. A further requirement is to ensure reproducibility by making data and methods transparent while minimizing exposure of personal information.

Several considerations are specific to our study. Because our work involves a Wikimedia project, we followed the guidelines outlined by Asikin-Garmager (2025) for conducting research on Wikimedia communities. Although the project did not require direct community interaction, we nevertheless informed the Wikidata community, incorporated their feedback, and contributed fixes to data issues, such as duplicate authors.[4] Second, while the demographic information we use from Wikidata–such as gender, citizenship, and institutional affiliations–is publicly available, we have taken precautions to avoid exposing individual researchers. We therefore share only a pseudonymized dataset that enables the recomputation of the diversity metrics without revealing personal details.

Finally, our use of the Genderize API to infer authors' gender and nationality carries known limitations. The tool excludes non-binary identities and has been reported to overcount men (Lockhart, King, and Munsch 2023). Following the principles outlined by Lockhart, King, and Munsch (2023), we rely only on aggregate demographic estimates derived from names. Furthermore, we acknowledge that relying on automated inference tools introduces additional sources of uncertainty that may systematically affect downstream analyses. Name-based gender inference is sensitive to cultural and linguistic variation (VanHelene et al. 2024), potentially leading to uneven error rates across regions and amplifying existing representation biases in the data. Moreover, inferred attributes are probabilistic and may obscure within-group heterogeneity, particularly when aggregated at scale. As shown in our dataset, a substantial share of demographic attributes is derived from such tools rather than directly observed, which may propagate their biases into diversity indicators. We therefore recommend that future work treats inferred attributes as estimates, reports their proportion explicitly, and, where possible, performs sensitivity analyses or compares results across subsets with and without inferred data. Despite these limitations, we consider such tools useful for statistical assessments of diversity patterns in scholarly publications.

Data Gathering

The following paragraphs describe our approach for analyzing Wikidata scientific articles, as summarize by Figure 1.

Wikidata We are interested in *scholarly works*, which we define as Q-items that contain the statement `instance of (P31) scholarly article (Q13442814)`, and *authors*, that is, Q-items that are referenced in a scholarly article statement via the property `author (P50)`.

While Wikidata contains more than 45M scholarly articles, an ideal publication record for our exploratory analysis should at least link the article QID to authors' QIDs. The availability of authors' QIDs is essential, since the author items can be processed further to extract gender, citizenship

[4] Deduplicate authors in P50 for scholarly articles in Wikidata, https://phabricator.wikimedia.org/T388968, accessed 2026-04-10.

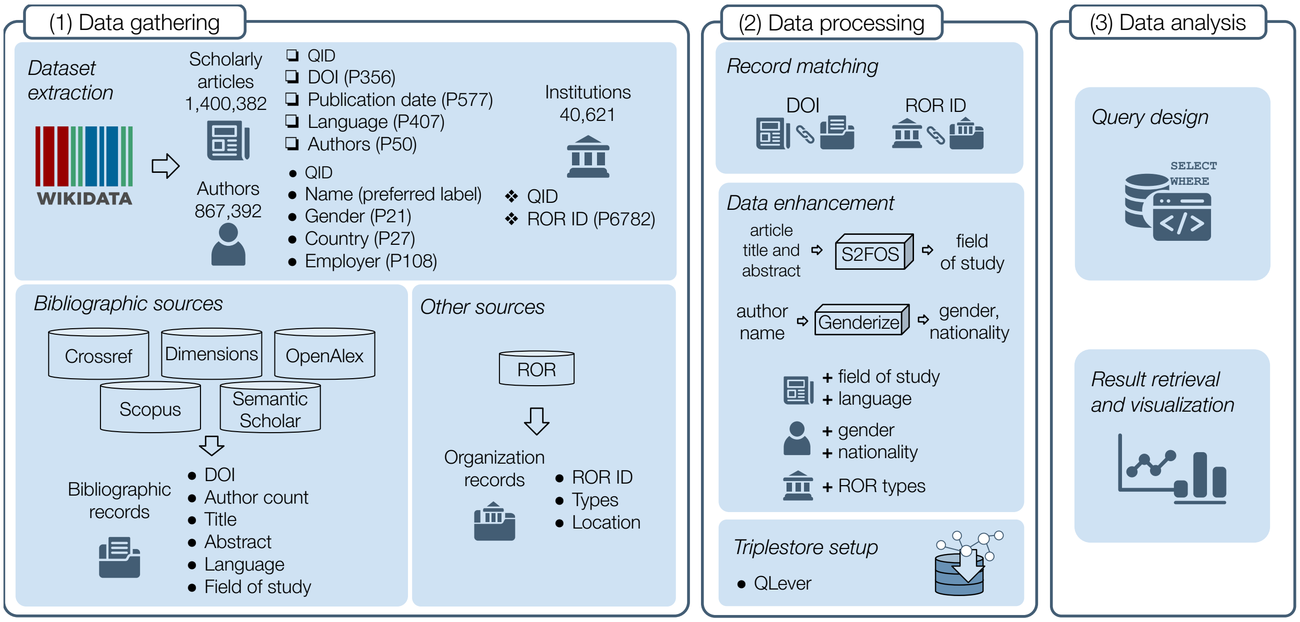

Figure 1: The *DivinWD* data processing pipeline.

and affiliation information. We also processed the data to filter out records that contained inconsistent data, such as multiple publication years for articles, and we filter out non-human authors[5]. The constraints we have applied in the article selection are the following:

1. The article item contains the statement `instance of (P31) scholarly article (Q13442814)`.
2. Each author must have an associated item and be human. In other words: (a) the publication must reference each author with the property `author (P50)`; (b) the publication must not contain the property `author name string (P2093)`, because its values are string literals that do not link to any item; (c) each author item must have a statement `instance of (P31) human (Q5)`.
3. The article must have a consistent publication year. Specifically, if multiple statements use the property `publication date (P577)`, all date values must fall within the same year.

As described in the following sections, to ensure that the list of authors for a given paper in Wikidata is complete and consistent, we cross-check it with the author list obtained from other bibliographic sources.

Scholarly Works. We examine three features related to scholarly works:

- *Publication year*, derived from property `publication date (P577)`. When finer temporal granularity is needed to validate time-dependent author attributes (nationality, affiliation), we select the earliest available date.
- *Language*, available through property `language of work or name (P407)`. We rely on language codes provided by Crossref and OpenAlex when Wikidata articles lack this piece of information.
- *Field of study (FOS)*, which refers to the branch of knowledge related to the subject of the article. We classify the field of study according to the set of 23 categories used by Semantic Scholar in the S2 FOS model[6], and either retrieve keywords from corresponding records in Semantic Scholar or compute labels from scratch using titles and abstracts extracted from Crossref, Dimensions, OpenAlex, and Scopus. This categorization ranges from hard sciences, such as physics, to soft ones, including philosophy.

Authors. The diversity of authors is analyzed considering the following aspects:

- *Gender*, through the property `sex or gender (P21)`.
- *Nationality*, through the property `country of citizenship (P27)`. Authors may have multiple countries of citizenship. Statements about nationality may be enriched with time qualifiers `start date (P580)` and `end date (P582)`; if present, they are compared with the publication date of the given article.
- *Affiliation*, which depends on the institutions to which the authors are employed at the time of publication of the articles. This information derives from property `employer (P108)`, whose values are classified according to the Research Organization Registry (ROR). Similarly to nationality, qualifiers `P580` and `P582` ensure the time compatibility of a given author's affiliation with respect to the publication date of their article.

[5] For example, some articles are collectively authored by collaborations (e.g. Q34344580) and there are also 8 pets appearing as authors (e.g. Q29597859).

[6] https://github.com/allenai/s2_fos, accessed 2026-04-10.

Unlike nationality, affiliations may vary more frequently over time; therefore, we exclude `P108` statements that lack both qualifiers. Authors may contribute to multiple institutional types, as they can be affiliated with different organizations simultaneously.

Whenever gender and nationality data are unavailable in Wikidata, we infer these attributes from author names using the Genderize API[7]. Author names are derived from preferred labels, which are associated with the corresponding entities in Wikidata.

For our analysis, we used a snapshot of Wikidata based on the RDF dump downloaded on 1 May 2025.

External sources This study compares Wikidata with other datasets for three main reasons: (1) we want to know whether the number of distinct authors provided by Wikidata for a given article agrees with one or more external sources; (2) we augment Wikidata by integrating missing data and reusing existing taxonomies; and (3) we provide insights into the coverage of Wikidata publications by other bibliographic databases.

We gathered structured data from Crossref, Dimensions, OpenAlex, Scopus, Semantic Scholar, and ROR. The corresponding datasets were collected on April 3, October 25, May 7, November 12, May 27, and May 20, 2025, respectively.

For each record indexed by Crossref, Dimensions, OpenAlex, Scopus, and Semantic Scholar, we retain the DOI, the title, and the number of distinct authors. Additionally, we collect language codes from Crossref and OpenAlex, and abstracts from Crossref, Dimensions, OpenAlex, and Scopus. From Semantic Scholar, we extract field-of-study information when available: each record may contain an array of fields of study, where each field has a label and a source indicating whether it originates from external databases or from Semantic Scholar's S2FOS prediction model. For our analysis, we retain the first S2FOS-predicted label from this array when present. From ROR records, we extract ROR IDs, types, and geographic locations.

Data Processing

To process Wikidata content, we utilize the N-Triples dump queried through a local QLever[8] triplestore.

Record Matching To establish correspondences between Wikidata and external datasets, we assign unique keys to each of their records. For publications, we use the DOI, while for institutions, we employ the ROR ID. Prior to the matching process, we exclude records lacking the necessary fields and those with either multiple or shared keys. Table 1 shows how each dataset compares to Wikidata in terms of records and distinct keys.

Data Enhancement. Firstly, we assign a single field of study to each article. For Wikidata publications with corresponding Semantic Scholar entries, we directly use the preexisting labels obtained during the external dataset preparation phase and the subsequent matching procedure. When Semantic Scholar labels are unavailable, we use titles and abstracts from Crossref, Dimensions, OpenAlex, and Scopus when present, and infer the field of study through S2FOS, excluding articles lacking abstracts to avoid potential misclassification due to title ambiguity. We always assign the prediction with the highest confidence score as the primary field of study for a given article.

Furthermore, articles lacking language information inherit language codes found in Crossref and OpenAlex.

Secondly, the Genderize API infers the missing values of gender and nationality of authors using names previously extracted from Wikidata.

Finally, we pair Wikidata organizations with their respective data gathered from ROR.

5 Dataset

Dataset. The *DivinWD* dataset is available on Zenodo at https://zenodo.org/records/18234750.[9] The source code for the data pipeline is available at https://github.com/DivinWD/dataset-resources and it is released under the MIT license.[10] Data and other content is released under the Creative Commons CC0 dedication (public domain).[11] The datasheet for the dataset is available at https://divinwd.dev/datasheet.

The dataset consists of a single GZip-compressed N-Triples file (`divinwd.nt.gz`) containing 23,486,895 RDF triples, i. e. the Wikidata subgraph of scholarly articles and authorship information explored in this study. In addition to the information extracted from Wikidata, the file includes custom triples, using the `divinwd.dev` base URI, which encode data derived from external sources, for which we defined new subjects, predicates, and objects to represent the additional attributes of articles, authors, and organizations. The full list of properties is shown in Table 2.

We pseudonymized the dataset by mapping Wikidata QIDs to a new set of pseudonyms (XIDs). Statement identifiers were pseudonymized as well since their original form encodes references to subject QIDs. In addition, labels were removed. ROR IDs and entity QIDs for languages, gender values, countries, continents, and useful constants–examples are `scholarly article (Q13442814)` and `human (Q5)`–were kept unaltered.

The code repository includes instructions to setup the pipeline, along with Python scripts that replicate the analysis of this paper. This repository also contains sample SPARQL queries that can be used as templates for constructing other queries against this dataset, along with details regarding the content and RDF format of the dataset file.

[7] https://genderize.io, accessed 2026-04-10.

[8] https://github.com/ad-freiburg/qlever, accessed 2026-04-10.

[9] The dataset has been assigned the DOI `DOI:10.5281/zenodo.18234750`

[10] The code release has been assigned the DOI `DOI:10.5281/zenodo.18267672`

[11] https://github.com/DivinWD/dataset-resources/blob/main/LICENSE.md

[12] https://creativecommons.org/publicdomain/zero/1.0/

Source	Records	Date	Unique records	Matching records	%
DivinWD (articles)	1,400,382	2025-05-01	1,209,332	-	100
Crossref	167,008,748	2025-04-03	167,008,744	1,183,029	97.8
Dimensions	1,163,716†	2025-10-25	1,162,280	1,162,280	96.1
OpenAlex	266,800,220	2025-05-07	173,229,377	1,188,395	98.3
Scopus	619,451†	2025-11-12	609,818	609,818	50.4
Semantic Scholar	226,689,623	2025-05-27	127,144,250	1,112,026	92.0
DivinWD (organizations)	40,621	2025-05-01	25,904	-	100
ROR	116,172	2025-05-20	116,172	25,904	100

Table 1: Records retrieved from each data source, unique keys (DOIs for articles, ROR IDs for organizations), records matched with Wikidata and percentage matching coverage. Sources marked with † (Dimensions and Scopus) were queried using DOIs extracted from Wikidata, therefore they are a subset of the Wikidata records.

6 Results

Statistics. The combined coverage of all bibliographic sources is substantial relative to our dataset: $1{,}191{,}752$ Wikidata scholarly articles (98.2%) have at least one correspondence, $1{,}183{,}202$ of which (99.2%) exhibit consistent author counts confirmed by at least one source. Regarding institutional affiliations, many Wikidata items were excluded due to missing ROR identifiers. Nevertheless, we successfully established correspondences for every item associated with a unique ROR ID.

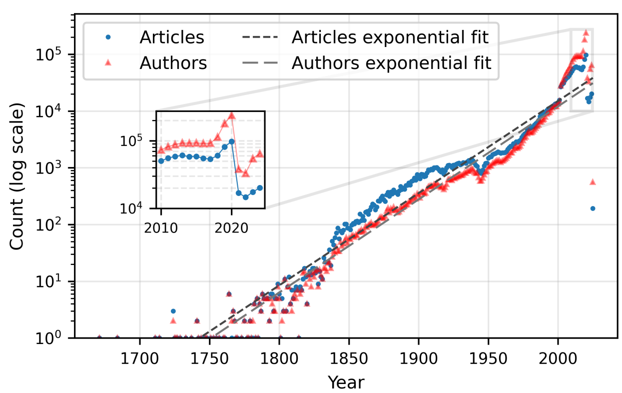

Figure 2: Annual article and author counts over time. Points show observed data, while dashed and solid lines represent exponential regression fits. Articles: $y = 4.7 \cdot 10^{-29} \cdot \exp(0.037 \cdot x)$, $R^2_{\text{articles}} = 0.94$, authors: $y = 1.9 \cdot 10^{-29} \cdot \exp(0.038 \cdot x)$, $R^2_{\text{authors}} = 0.95$.

Articles. The full set of scholarly articles on Wikidata includes articles published from 1671 to 2025. Figure 2 shows a log-linear plot of published articles and authors for each year. The overall trends are exponential and we observe a peculiar fluctuation in the time window of 2010-2024, highlighted in the plot, during which the number of publications peaks in 2020, then it drops dramatically in the following years. This behavior stems from a temporary halt in large-scale scholarly data imports into Wikidata, prompted by the project approaching the limits of its infrastructure.[12] By 2025, the Wikidata Query Service (WDQS) contained over 16 billion RDF triples and was growing by roughly 1 billion triples per year. This growth caused several problems to the overall Wikidata infrastructure and Wikidata's infrastructure could not accommodate mass imports. As a result, the community adopted two measures: first, it suspended further imports of scholarly datasets for the near future; second, on 9 May 2025, the WDQS was split into two graphs–`main` and `scholarly`–moving all scholarly entities to a separate Blazegraph instance.[13] We will focus our analysis in the time window 2010-2024, which comprises $758{,}991$ publications and $766{,}572$ authors in total.

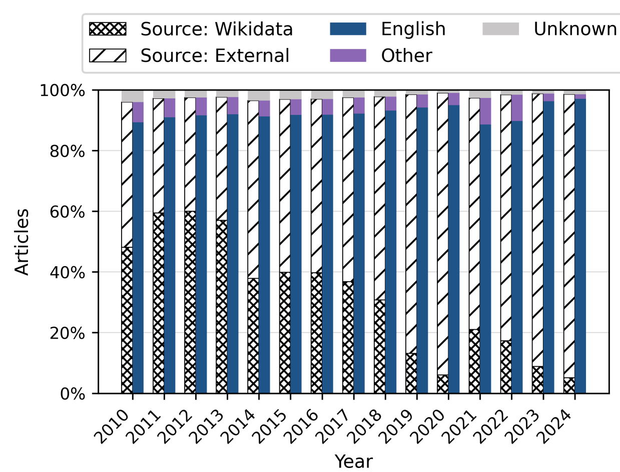

Figure 3: Distribution of articles' data source and language by year.

Language. Figure 3 shows the distribution of articles by language. For each year, the plot shows two bars showing the data source and the language (English or other). After 2013, most of the languages are derived from Crossref and

[12]https://www.wikidata.org/wiki/Wikidata:WikiProject_Limits_of_Wikidata

[13]https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_graph_split

OpenAlex. Non-English scholarly publications are predominantly written in Chinese, and European languages such as French, German, Spanish, and Italian.

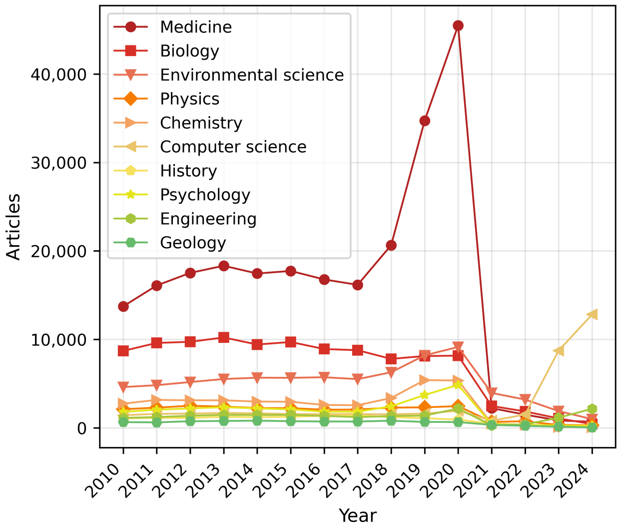

Figure 4: Top ten fields of study throughout 2010-2024.

Field of Study. The ten most frequent fields of study across the entire Wikidata corpus–consisting of 1,120,505 articles published between 1671 and 2025–are Medicine and the hard sciences, with History and Psychology as notable exceptions. Figure 4 presents yearly publication counts from 2010 to 2024, showing that medicine remains the most represented field until 2020. In contrast, computer science–and engineering to a lesser extent–exhibit an increase in the last two years, likely reflecting the recent surge of publications related to artificial intelligence.

Authors. In the results that follow, we focus on authors who published at least one article between 2010 and 2024. Those authors are counted at most once each year, regardless of the number of publications to which they contributed.
Gender. For each year in Figure 5, the bar on the left compares authors whose gender was extracted from Wikidata with those whose gender was inferred from names using the Genderize service. Males are the most represented category (70%), with the share of female authors slightly increasing between 2010 and 2020. In the plot, *Other* is a broader category that includes authors with non-binary gender values in Wikidata, we report the full table on non-binary genders available in Wikidata in Appendix B. Overall, non-binary authors are less than 0.10% for every publication year considered. There are 6,052 authors (0.79%) whose gender is unknown.
Country of citizenship. We analyze authors' nationalities using the `country of citizenship (P27)` attribute when available. Figure 6 aggregates countries by continent and shows that European authors are the most represented overall. At the country level, however, the most prominent contributors are United States, China, Italy, Germany and Poland. The nationality of 22,325 authors (2.9%) could not be determined. When available, country information was primarily inferred via Genderize (88% of cases), while only a small fraction of authors had explicit nationality data recorded in Wikidata.

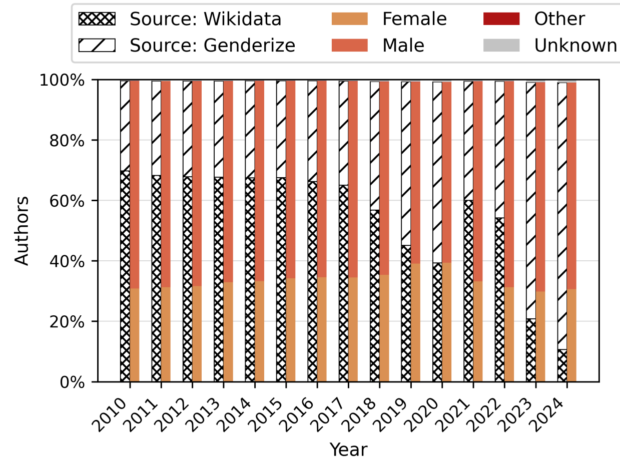

Figure 5: Distribution of author genders by year. The bars on the left show the source of the data.

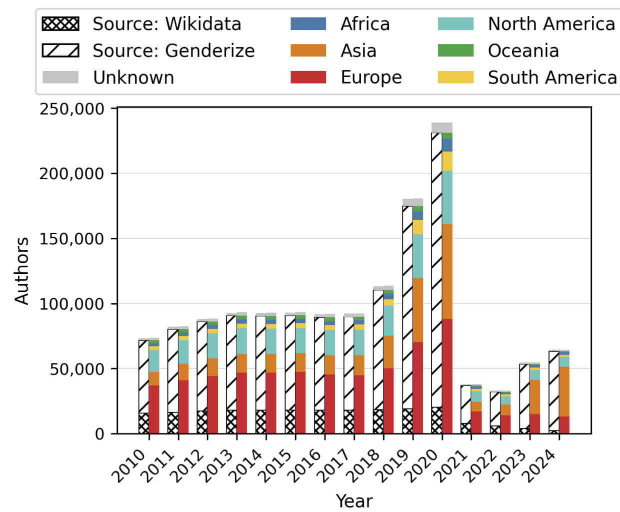

Figure 6: Distribution of authors by continent and by year. The bars on the left show the source of the data.

Figure 7 presents a heatmap illustrating the relationship between authors' country of citizenship (rows) and the country hosting their affiliation institution (columns), aggregated at the continental level. The figure suggests that European and North American institutions attract scientists from all regions. With respect to outward mobility, European scientists tend to remain in Europe or move to North America, while North American scientists more often relocate to Europe or Oceania. In Appendix C we report more detailed data at the European level.
Affiliation. Trends in the affiliation dimension are presented in Figure 8. The classification is taken from ROR and it is

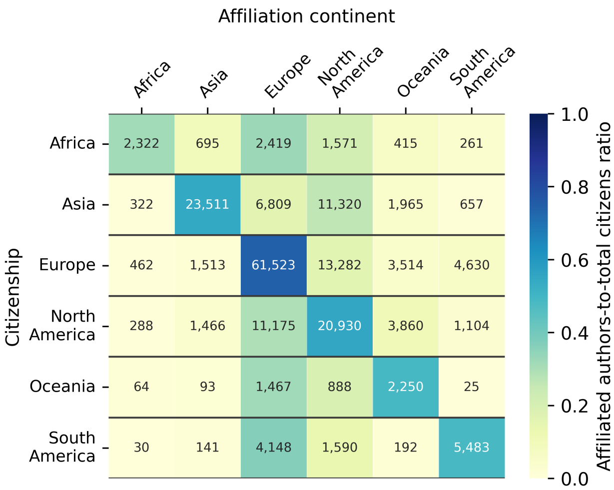

Figure 7: Heatmap showing authors' country of citizenship (rows) and the countries where their affiliation institution is hosted (columns), aggregated at the continental level. Cell values indicate the number of authors, while color intensity reflects the ratio of affiliated authors relative to the total number of authors for each citizenship group (row).

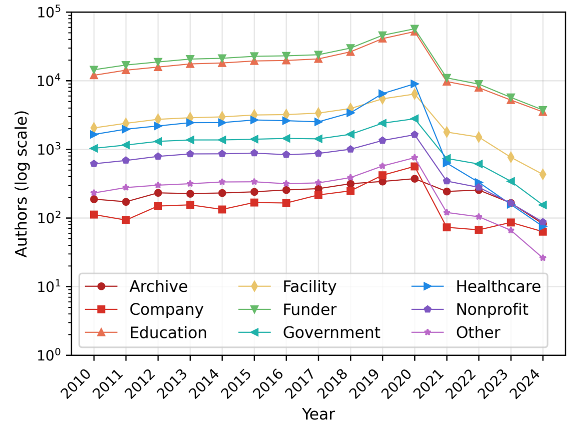

Figure 8: Number of authors by affiliation category between 2010 and 2024, shown on a logarithmic scale.

explained in Appendix B. The *funder* and *education* types are the most common; their similarities suggest a correlation between academia and funding institutions, as is usually the case. Results concerning this dimension of diversity are limited: only 192,133 authors (25%) are associated with at least one affiliation type.

7 Discussion

This work highlights several lessons learned from exploring Wikidata as a source for computing diversity indicators in scientific publications.

- **Scope matters for diversity monitoring.** While a global perspective on diversity in science is valuable, our analysis suggests that diversity monitoring efforts benefit from a more narrowly defined scope. Restricting the scope can mitigate a range of data-related challenges, including inconsistencies in historical data (e. g. citizenship of countries that don't exist anymore), heterogeneous publication practices across disciplines (e.g., journals versus conferences), and uneven data completeness. A focused scope also enables closer engagement with specific scientific communities, facilitates the application of domain expertise, and supports more manageable data curation and validation processes.
- **Biases are pervasive and cumulative.** Large-scale bibliographic datasets inevitably reflect biases originating from multiple sources. In our case, some biases stem from the uneven coverage of scholarly data in Wikidata, which is particularly visible after the import freeze enacted around 2020. Other biases may be inherited from external datasets that were imported into Wikidata. For instance, the over-representation of European authors in our corpus may reflect broader systemic biases in the Wikimedia ecosystem or the greater availability of structured data about European scientists in source databases. These factors complicate the interpretation of diversity indicators and call for caution when drawing comparative conclusions.
- **Automated inference tools require careful interpretation.** Automated methods for inferring demographic attributes introduce additional sources of uncertainty and bias. Tools such as Genderize, for example, systematically exclude non-binary gender identities and exhibit known limitations when applied to names from certain linguistic or cultural contexts. While such tools can support large-scale analyses, their outputs should be interpreted as probabilistic estimates rather than ground truth.
- **Privacy considerations are central to diversity analyses.** Diversity monitoring initiatives must balance transparency and accountability with the protection of individual researchers' privacy. Although Wikidata enables the construction of rich, person-level bibliometric datasets, many of the attributes used in this study–such as gender and nationality–are sensitive and may be misused. To mitigate these risks, we adopted a conservative data-sharing strategy: the released dataset is anonymized, results are reported exclusively in aggregated form, and no information enabling the identification of individual authors is disclosed, even if the original data is publicly available. We argue that this approach provides meaningful insights into the *status quo* of diversity in scientific publishing while respecting the rights and dignity of individual researchers.

Taken together, these limitations do not diminish the value of this study. Rather, they underscore both the challenges and the potential of using Wikidata as an infrastructure for monitoring diversity in science. Our results demonstrate that, despite current shortcomings, Wikidata can serve as a flexible and extensible foundation for computing diversity indicators, particularly when adopted by well-defined scientific communities. By curating conference series or disciplinary corpora within Wikidata, communities can leverage existing tools to compute diversity metrics with relatively low overhead. Moreover, as a continuously evolving, community-maintained knowledge graph, Wikidata is likely to improve in coverage and quality over time. We view this work as a contribution to that broader open infrastructure and as a step toward more transparent and inclusive assessments of scientific diversity.

8 Conclusions

In this paper we introduced the *DivinWD* dataset for assessing diversity in scholarly publications represented in Wikidata. By combining Wikidata's openly available knowledge graph with five major bibliographic datasets–Crossref, Dimensions, OpenAlex, Scopus and Semantic Scholar–we compiled a harmonized corpus of roughly 1.2 million articles and published between 2010 and 2024. We augmented the dataset with the gender and nationality of researchers. The dataset comprises more than 23 million RDF triples and we also release the pipeline under an open, permissive license.

We leveraged this dataset to analyze the diversity of scholarly publications in Wikidata. Our exploration uncovered pronounced coverage biases: English remains the overwhelmingly dominant language and men account for about 70% of the author population. Non-binary genders comprise less than 0.1% of authors, reflecting limitations both in Wikidata's coverage and in the binary assumptions built into external inference tools.

Our findings underscore both the promise and the current limitations of using Wikidata to measure diversity in science; addressing these limitations will be crucial to ensure that open knowledge graphs truly reflect the plurality of global scholarship.

Acknowledgments

The authors gratefully acknowledge the "Scientometric Researcher Access to Data" (SRAD) program by Digital Science for providing free-of-charge access to Dimensions (https://www.dimensions.ai) and Demografix ApS for providing free-of-charge access to the Genderize API (https://genderize.io). The authors are also grateful to Daniel Mietchen, Camillo Pellizzari, James Hare, Pablo Aragón, and Dani Metilli for their valuable feedback on an early version of this paper.

References

Adams, J. 2013. The fourth age of research. *Nature*, 497(7451): 557–560.

AlShebli, B. K.; Rahwan, T.; and Woon, W. L. 2018. The preeminence of ethnic diversity in scientific collaboration. *Nat. Commun.*, 9(1): 5163.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica.

Antin, J.; Yee, R.; Cheshire, C.; and Nov, O. 2011. Gender differences in Wikipedia editing. In *Proc. WikiSym*, 11–14.

Asikin-Garmager, E. 2025. Research and Privacy on Wikipedia. OSF Preprint.

Conroy, M. 2023. Quantifying the gap: the gender gap in French writers' Wikidata. *J. Cult. Anal.*, 8(2).

Dijksterhuis, W. P.; Stroes, C. I.; Tan, W.-L.; Ithimakin, S.; Calles, A.; van Oijen, M. G.; Verhoeven, R. H.; Barriuso, J.; Oosting, S. F.; Ivankovic, D. K.; et al. 2020. From presentation to paper: gender disparities in oncological research. *Int. J. Cancer*, 146(11): 3011–3021.

Dutt, K. 2020. Race and racism in the geosciences. *Nat. Geosci.*, 13(1): 2–3.

Erxleben, F.; Günther, M.; Krötzsch, M.; Mendez, J.; and Vrandečić, D. 2014. Introducing Wikidata to the linked data web. In *ISWC*, 50–65.

European Union. 2012. Charter of Fundamental Rights of the European Union. Official Journal of the European Union, C 326/391.

Freeman, R. B.; and Huang, W. 2014. Collaboration: Strength in diversity. *Nature*, 513(7518): 305–305.

Freeman, R. B.; and Huang, W. 2015. Collaborating with people like me: Ethnic coauthorship within the United States. *J. Labor Econ.*, 33(S1): S289–S318.

Freire, A.; Porcaro, L.; and Gómez, E. 2021. Measuring diversity of artificial intelligence conferences. In *AIDBEI*, 39–50.

Gómez, E.; Porcaro, L.; Frau, P.; and Vinagre, J. 2024. Diversity in Artificial Intelligence Conferences. Technical report, European Commission, Joint Research Centre. JRC137550.

HLEGAI. 2019. Ethics Guidelines for Trustworthy AI. Technical report, European Commission.

Hupont, I.; Gomez, E.; Tolan, S.; Porcaro, L.; and Freire, A. 2022. Monitoring diversity of AI conferences: Lessons learnt and future challenges in the DivinAI project. *arXiv*, arXiv:2203.01657.

Hupont, I.; Tolan, S.; Frau, P.; Porcaro, L.; and Gomez, E. 2024. Measuring and Fostering Diversity in Affective Computing Research. *IEEE Trans. Affect. Comput.*, 15(1): 63–78.

Hupont, I.; Tolan, S.; Freire, A.; Porcaro, L.; Estevez, S.; and Gómez, E. 2021. How diverse is the ACII community? Analysing gender, geographical and business diversity of Affective Computing research. In *Proc. ACII*, 1–8.

King, L.; MacKenzie, L.; Tadaki, M.; Cannon, S.; McFarlane, K.; Reid, D.; and Koppes, M. 2018. Diversity in geoscience: Participation, behaviour, and the division of scientific labour at a Canadian geoscience conference. *Facets*, 3(1): 415–440.

Klein, M.; Gupta, H.; Rai, V.; Konieczny, P.; and Zhu, H. 2016. Monitoring the Gender Gap with Wikidata Human Gender Indicators. In *Proc. OpenSym*.

Konieczny, P.; and Klein, M. 2018. Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society*, 20(12): 4608–4633.

Lockhart, J. W.; King, M. M.; and Munsch, C. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nat. Hum. Behav.*, 7(7): 1084–1095.

Melis, B.; Fioravanti, M.; Paolini, C.; and Metilli, D. 2025. How have you modelled my gender? Reconstructing the history of gender representation in Wikidata. *Internet Histories*, 9(1–2): 163–179.

Melis, B.; Paolini, C.; Fioravanti, M.; and Metilli, D. 2024. What does it mean to be queer in Wikidata? Practices of gender representation within a transnational online community. *Commun. Cult. Crit.*, 17(3): 200–207.

Metilli, D.; and Paolini, C. 2023. Non-binary Gender Representation in Wikidata. In *Ethics in Linked Data*. Litwin Books.

Mora-Cantallops, M.; et al. 2019. A systematic literature review on Wikidata. *Data Technol. Appl.*, 53(3): 250–268.

Nielsen, F. Å.; Mietchen, D.; and Willighagen, E. 2017. Scholia, scientometrics and Wikidata. In *ESWC*, 237–259.

Oswald, A.-M. M.; and Ostojic, S. 2020. Curating more diverse scientific conferences. *Nat. Rev. Neurosci.*, 21(11): 589–590.

Partiali, B.; Oska, S.; Touriel, R. B.; Delise, A.; Barbat, A.; and Folbe, A. 2021. Gender disparity in speakers at a major academic emergency medicine conference. *Emerg. Med. J.*, 38(5): 379–380.

Patel, H.; Chen, T.; Bongiovanni, I.; and Demartini, G. 2024. Estimating Gender Completeness in Wikipedia. *arXiv*, arXiv:2401.08993.

Pfeil, U.; Zaphiris, P.; and Ang, C. S. 2006. Cultural differences in collaborative authoring of Wikipedia. *J. Comput.-Mediat. Commun.*, 12(1): 88–113.

Piscopo, A.; and Simperl, E. 2019. What we talk about when we talk about Wikidata quality: a literature survey. In *Proc. OpenSym*.

Porcaro, L.; Vinagre, J.; Frau, P.; Hupont, I.; and Gómez, E. 2023. Behind Recommender Systems: the Geography of the ACM RecSys Community. *arXiv*, arXiv:2309.03512.

Rossi, D.; and Zacchiroli, S. 2022. Geographic diversity in public code contributions: an exploratory large-scale study over 50 years. In *Proc. MSR*, 80–85.

Shishkova, E.; Kwiecien, N. W.; Hebert, A. S.; Westphall, M. S.; Prenni, J. E.; and Coon, J. J. 2017. Gender diversity in a STEM subfield–analysis of a large scientific society and its annual conferences. *J. Am. Soc. Mass Spectrom.*, 28(12): 2523–2531.

Small, Z. 2023. Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History. New York Times.

Spector, N. D.; Asante, P. A.; Marcelin, J. R.; Poorman, J. A.; Larson, A. R.; Salles, A.; Oxentenko, A. S.; and Silver, J. K. 2019. Women in pediatrics: progress, barriers, and opportunities for equity, diversity, and inclusion. *Pediatrics*, 144(5): e20192149.

Swartz, T. H.; Palermo, A.-G. S.; Masur, S. K.; and Aberg, J. A. 2019. The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. *J. Infect. Dis.*, 220(Supplement_2): S33–S41.

Trinkenreich, B.; Wiese, I.; Sarma, A.; Gerosa, M.; and Steinmacher, I. 2022. Women's Participation in Open Source Software: A Survey of the Literature. *ACM Trans. Softw. Eng. Methodol.*, 31(4).

United Nations. 1948. Universal Declaration of Human Rights. General Assembly Resolution 217 A (III).

VanHelene, A. D.; Khatri, I.; Hilton, C. B.; Mishra, S.; Gamsiz Uzun, E. D.; and Warner, J. L. 2024. Inferring gender from first names: Comparing the accuracy of Genderize, Gender API, and the gender R package on authors of diverse nationality. *PLOS Digit. Health*, 3(10): e0000456.

Vila-Concejo, A.; Gallop, S. L.; Hamylton, S. M.; Esteves, L. S.; Bryan, K. R.; Delgado-Fernandez, I.; Guisado-Pintado, E.; Joshi, S.; da Silva, G. M.; Ruiz de Alegria-Arzaburu, A.; et al. 2018. Steps to improve gender diversity in coastal geoscience and engineering. *Palgrave Commun.*, 4(1): 1–9.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proc. ICWSM*, volume 9, 454–463.

Wagner, C.; Graells-Garrido, E.; Garcia, D.; and Menczer, F. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Sci.*, 5(1): 5.

Willighagen, E.; Mietchen, D.; Patel-Schneider, P.; Linden, K.; Willighagen, L.; and Fahl, W. 2026. Scholia 2026: Compliance with SPARQL 1.1.

Yang, Y.; Tian, T. Y.; Woodruff, T. K.; Jones, B. F.; and Uzzi, B. 2022. Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proc. Natl. Acad. Sci. U.S.A.*, 119(36): e2200841119.

Yunus, Y.; Chen, T.; and Demartini, G. 2025. Exploring Wikipedia Gender Diversity Over Time – The Wikipedia Gender Dashboard (WGD). *arXiv*, arXiv:2501.12610.

Zhang, C. C.; and Terveen, L. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *Proc. OpenSym*.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, please refer to Sub-section *"Responsible Data Collection Practices"* in Section 4.
   (g) Did you discuss any potential misuse of your work? Yes, please refer to Sub-section *"Responsible Data Collection Practices"* in Section 4.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we pseudonymized all personal data about authors. Yes, please refer to Sub-section *"Responsible Data Collection Practices"* in Section 4. We adhere to the ethical guidelines about research on Wikipedia and privacy as outlined by Asikin-Garmager (2025).
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A, we did not train new models, we relied on a pre-trained model and an external API: (1) for the computation of the "Field of Study," we used the `s2_fos`[1] model from Allen AI; and (2) for determining the gender and nationality of authors we used the Genderize API.[2]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A, we did not train new models.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A, we did not train new models.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we discuss the biases that the automatic classification of gender and nationality can introduce in Section 5 of the paper.
3. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we used public and commercially-available bibliographic datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, to perform this study we handled personal data. Please refer to Sub-section *"Responsible Data Collection Practices"* in Section 4. We processed only metadata about scientific publication and we did not process any paper content.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. Our project is developed to be part of the Wikimedia ecosystem, which has the principles of openness and transparency among its core inspiring values, and as such it strictly adheres to these principles. We are following the FAIR principles:
      - Findability and Accessibility: all the code and data generated in the project are released under an open source, permissive license, and made available on public repositories. We archived the code on Zenodo to obtain a permanent DOI for the code release to improve findability.
      - Interoperability and Reusability: the dataset is released in a standard format (N-Triples).
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes
4. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA, the study does not involve human participants or crowdsourcing.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA, the study does not involve human participants.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA, no participants were recruited.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes

[1] https://github.com/allenai/s2_fos, accessed on 2026-01-16.
[2] https://genderize.io/, accessed on 2026-01-16.

Generative AI Usage Statement

We confirm that all text in this paper was written by the authors. In preparing this paper, we utilized ChatGPT 5.2 (Pro version) for grammar and spelling checks and to improve the clarity of the author-written text. No Generative AI tools were used to generate content and citations from scratch. The content and intellectual contributions remain entirely those of the human authors. We acknowledge the contributions of the mentioned tool in enhancing the writing process while maintaining full academic integrity and abiding by the AAAI policy.

References

Asikin-Garmager, E. 2025. Research and Privacy on Wikipedia. OSF Preprint. https://osf.io/preprints/osf/uyxnf_v2.

A Additional Bibliographic Data Sources

Crossref.[14] Crossref is a nonprofit organization and the largest Digital Object Identifier (DOI) registration agency in the world. Crossref was founded in 1999 and enables persistent citation linking by assigning DOIs to research outputs and ensuring metadata interoperability. Unlike other bibliographic databases it does not host full-text content but provides infrastructure to interlink distributed scholarly resources, making metadata openly available.

Dimensions.[15] Dimensions was launched in 2018 by Digital Science, a technology company that maintains several products and solutions for researchers. Dimensions offersadditional data including grants, clinical trials, patents, and policy documents. While they offer commercial access to their data, they provide it free of charge for bibliometric scientific research.

OpenAlex.[16] OpenAlex is an open-access bibliographic database lauched in 2022 that provides comprehensive metadata on scholarly works, authors, and institutions as a free alternative to commercial platforms. OpenAlex focuses its coverage on diamond open-access journals which are often underrepresented elsewhere and it offers an API for large-scale retrieval to support research evaluation and open science.

Scopus.[17] Scopus is a scientific database launched in 2004 by Elsevier, accessible via subscription. As of 2025, Scopus' database consists of more than 100 million records, spanning articles, books, conference proceedings, preprints, and patents. Scopus is widely used for academic research and research evaluation and covers publications from 1788 to the present.

Semantic Scholar.[18] Semantic Scholar (S2) is a free academic search engine developed by the Allen Institute for AI.[19] Semantic Scholar was launched in 2015 and leverages natural language processing and machine learning to extract structured metadata, identify key concepts, and map citation relationships, thereby enabling advanced discovery and analysis of scientific literature.

Research Organization Registry (ROR).[20] The Research Organization Registry is a community-led, open-access database that provides persistent identifiers for research organizations globally. ROR was launched in 2019 by the California Digital Library, Crossref, and DataCite and it aims to standardize the identification of institutions and make them freely available under a free CC0 license. The registry is maintained through a collaborative curation process and is updated monthly.

B Additional Information

Full list of properties using in *DivinWD*. The full list of properties in the *DivinWD* dataset is shown in Table 2.

[14]https://www.crossref.org/, accessed 2026-06-01.
[15]https://www.dimensions.ai, accessed 2026-06-01.
[16]https://openalex.org/, accessed 2026-06-01.
[17]https://www.scopus.com/, accessed 2026-06-01.
[18]https://www.semanticscholar.org/, accessed 2026-06-01.
[19]https://allenai.org/, accessed 2026-06-01.
[20]https://ror.org/, accessed 2026-06-01.

Property (`prefix`)	Count
http://www.w3.org/1999/02/22-rdf-syntax-ns#type (`rdf:type`)	3,985,655
http://www.w3.org/2000/01/rdf-schema#label (`rdfs:label`)	3,873
http://www.wikidata.org/prop/direct/P21 (`wdt:P21`)	379,798
http://www.wikidata.org/prop/direct/P297 (`wdt:P297`)	229
http://www.wikidata.org/prop/direct/P30 (`wdt:P30`)	2,220
http://www.wikidata.org/prop/direct/P31 (`wdt:P31`)	2,406,883
http://www.wikidata.org/prop/direct/P407 (`wdt:P407`)	568,526
http://www.wikidata.org/prop/direct/P50 (`wdt:P50`)	5,151,455
http://www.wikidata.org/prop/direct/P571 (`wdt:P571`)	32,881
http://www.wikidata.org/prop/direct/P576 (`wdt:P576`)	3,288
http://www.wikidata.org/prop/direct/P577 (`wdt:P577`)	1,446,737
http://www.wikidata.org/prop/direct/P6782 (`wdt:P6782`)	26,078
http://www.wikidata.org/prop/P108 (`p:P108`)	721,513
http://www.wikidata.org/prop/statement/P108 (`ps:P108`)	721,434
http://www.wikidata.org/prop/P27 (`p:P27`)	109,743
http://www.wikidata.org/prop/statement/P27 (`ps:P27`)	109,679
http://www.wikidata.org/prop/P463 (`p:P463`)	76,629
http://www.wikidata.org/prop/statement/P463 (`ps:P463`)	6,066
http://www.wikidata.org/prop/qualifier/P580 (`pq:P580`)	538,376
http://www.wikidata.org/prop/qualifier/P582 (`pq:P582`)	273,461
https://divinwd.dev/genderize/gender (`genderize:gender`)	2,151,238
https://divinwd.dev/genderize/nationality (`genderize:nationality`)	2,099,281
https://divinwd.dev/oacr/lang (`oacr:lang`)	1,185,545
https://divinwd.dev/ror/id (`ror:id`)	116,172
https://divinwd.dev/ror/location (`ror:location`)	116,046
https://divinwd.dev/ror/type (`ror:type`)	133,584
https://divinwd.dev/semanticscholar/fos/prediction (`ss:fos-prediction`)	78,901
https://divinwd.dev/semanticscholar/fos/value (`ss:fos-value`)	1,041,604

Table 2: Distribution of RDF properties in the *DivinWD* dataset.

Non-Binary Gender Values in Wikidata. Table 3 lists the Wikidata items used to identify gender identities in the dataset, reporting their corresponding QIDs and English labels. All these items have been aggregated in our analysis under "Other." For a dedicated analysis, refer to the *WiGeDi* (Wikidata Gender Diversity) project.[21]

ROR organization categories. ROR describes organizations by defining a schema consisting of 11 fields;[22] specifically, the *type* field can be populated with one or more values from the following list of categories:

- Education: a university or similar institution involved in providing education and educating or employing researchers.

[21]https://wigedi.com, accessed 2026-04-10.
[22]https://ror.readme.io/docs/ror-data-structure, accessed 2026-04-10.

QID	Label (en)
Q505371	agender
Q18116794	genderfluid
Q12964198	genderqueer
Q121307094	intersex man
Q121307100	intersex woman
Q48270	non-binary
Q130477254	non-binary woman
Q27679684	transfeminine
Q189125	transgender
Q27679766	transmasculine
Q17148251	travesti
Q2449503	trans man
Q1052281	trans woman
Q113124952	undisclosed gender

Table 3: Non-binary genders values in Wikidata.

- Funder: an organization that awards research funds or provides in-kind support.
- Healthcare: a medical care facility such as hospital or medical clinic. Excludes medical schools (categorized as "Education").
- Company: a private for-profit corporate entity involved in conducting or sponsoring research.
- Archive: an organization involved in stewarding research and cultural heritage materials. Includes libraries, museums, and zoos.
- Nonprofit: a non-profit and non-governmental organization involved in conducting or funding research.
- Government: an organization that is part of or operated by a national or regional government and that conducts or supports research.
- Facility: a specialized facility where research takes place, such as a laboratory, telescope or dedicated research area.
- Other: a category for any organization that does not fit the categories above.

C Additional Results

Country of citizenship. Figure 10 shows a heatmap of the relationship between authors' country of citizenship (rows) and the country where their affiliation institution is located (columns), restricted to European countries. Cell colors represent the share of affiliated authors relative to the total number of authors from each citizenship, while numbers indicate absolute counts. The diagonal highlights authors affiliated in their country of citizenship, and off-diagonal cells capture cross-border affiliations within Europe and beyond.

Diversity Indices. We present here the same diversity metrics used by the the DivinAI project, computed over the *DivinWD* dataset. We faced some additional challenges because, across multiple dimensions, individuals may belong to overlapping categories. We employ the Shannon diversity index H', defined as:

$$H' = -\sum_{i \in S} p_i \ln p_i \tag{1}$$

where S is the set of classes, and p_i is the proportional abundance of individuals belonging to class i in the population. H' captures both *richness*, i.e., the number of distinct classes present in the population, and *evenness*, that is, how uniformly individuals are distributed across classes.

The formulation above assumes mutually exclusive categories and limits the representation of diversity in the citizenship and affiliation dimensions, in which authors may belong to multiple classes simultaneously. Therefore, we propose an extension of Equation 1 by redefining p_i as follows: for each individual a in the population A, let M_a be the set of classes to which a belongs, where $1 \leq |M_a| \leq |S|$. Then:

$$p_i = \frac{1}{|A|} \sum_{a \in A} \frac{\mathbb{1}_{i \in M_a}}{|M_a|} \quad \forall i \in S \tag{2}$$

As an example, if author c is both Spanish and Swiss, then c contributes $1/2$ to each country. Figure 9 provides a comprehensive view of the diversity indices computed between 2010 and 2024 for all dimensions.

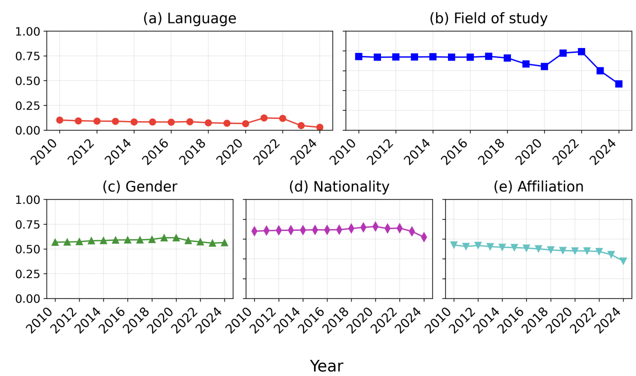

Figure 9: Shannon indices for language (a) and field of study (b) of articles, and for gender (c), country (d), and affiliation (e). The latter does not include type 'funder'. Each index is normalized in the range $[0, 1]$ by dividing H' by its maximum value $\ln |S|$. Specifically, the total number of distinct languages observed was $|S| = 54$; the fields of study are $|S| = 23$; the gender categories are $|S| = 3$; the observed countries were $|S| = 225$; concerning affiliation, we follow DivinAI taxonomies and do not consider the funder type, whose trend is highly similar to that of education: $|S| = 8$.

Figure 10: Heatmap of authors' country of citizenship (rows) versus country of affiliation (columns) for EU countries. Cell values indicate the number of authors, while color intensity reflects the ratio of affiliated authors relative to the total number of authors for each citizenship group (row).