Wiki Workshop (13th edition) – March 25-26, 2026

DivinWD: Insights from Wikidata for Measuring Diversity in Scholarly Publications

Zeno Saletti
DISI, University of Trento

Cristian Consonni[*,†]
European Commission
Joint Research Centre (JRC)

Pedro Frau[*]
European Commission
Joint Research Centre (JRC)

Emilia Gómez[*]
European Commission
Joint Research Centre (JRC)

Abstract

This paper introduces *DivinWD*, a curated dataset and reproducible pipeline that integrate Wikidata with multiple bibliographic sources to analyze diversity dimensions in scholarly publications, revealing systematic coverage biases and infrastructural constraints relevant to Wikimedia-based research and community-driven knowledge curation.

1 Introduction

The open science movement has underscored the importance of transparent and inclusive bibliographic infrastructures. As artificial intelligence increasingly mediates access to knowledge, the quality and openness of underlying data sources are critical for ensuring equitable and trustworthy information ecosystems. Diversity and representation in science are particularly salient: evidence shows that heterogeneous teams produce more novel and highly cited research, yet persistent inequalities still shape scientific participation. Monitoring diversity in the scientific record therefore supports both evaluation of diversity, equity and inclusion (DEI) initiatives and the design of fair AI systems. Initiatives such as *WikiCite*,[1] together with tools like Scholia[2] (Willighagen et al., 2026; Rasberry et al., 2019; Nielsen et al., 2017) and large-scale data imports, have positioned Wikidata as a promising foundation for open bibliographic repositories. In this paper, we present *DivinWD* (*DIVersity IN WikiData*) 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: how complete and accurate is the Wikidata coverage compared to established bibliographic infrastructures? What kinds of biases emerge when Wikidata is used as a lens for studying the diversity of scientific publications?

[*] *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.

[†] Corresponding author: `cristian.consonni@acm.org`

[1] `http://wikicite.org/`, accessed 2026-03-03.

[2] `https://scholia.toolforge.org/`, accessed 2026-03-03.

2 Methods

We begin with the May 2025 Wikidata dump, identifying items of type *scholarly article* via property `instance of` (P31). To ensure that diversity indicators can be computed at the author level, we retain only those articles whose authors are linked to Wikidata items via `author` (P50) statements and whose authors are instances of *human* (Q5). Articles that instead use the string-valued property `author name string` (P2093), or that list non-human entities, are excluded. After filtering for unique publication years and removing items with inconsistent metadata we obtain a dataset of 1.4 million articles, of which roughly 760 000 fall within the 2010–2024 period and have a complete set of author QIDs.

To evaluate the coverage and accuracy of Wikidata we cross-match the selected articles with other bibliographic databases. For each article we extract its digital object identifier (DOI), using this as key to locate corresponding records in Crossref, OpenAlex, Dimensions, Scopus and Semantic Scholar. We also ingest institutional records from the Research Organisation Registry (ROR) to classify author affiliations. Table 1 presents some statistics about our source datasets. Matching reveals that approximately 1,191,752 Wikidata articles (about 98.2%) have at least one counterpart in the external sources; among these, 99.2% exhibit author counts consistent with at least one external database, lending confidence to the completeness of the retained author lists.

2.1 Metadata extraction and enhancement

For each article in the matched dataset we extract the publication language (property `language of work` P407), the publication date (P577) and the DOI (P356). We assign a single field of study by combining existing Semantic Scholar classifications with predicted labels derived from titles and abstracts obtained through Crossref, OpenAlex, Dimensions and Scopus using the S2 FOS model[3]. Author items provide demographic attributes: gender is taken from `sex or gender` (P21) and grouped into female, male, other–which includes non-binary genders–

[3] `https://github.com/allenai/s2_fos`, accessed 2026-03-03.

and undetermined. We also augment the data using the Genderize API[4]. Nationality is drawn from `country of citizenship` (P27), accounting for time qualifiers and allowing authors to contribute fractional counts when they hold multiple citizenships. Institutional affiliation is derived from `employer` (P108), with qualifiers ensuring that employment dates overlap the publication year. We map institutional QIDs to ROR identifiers and apply ROR's taxonomy of nine organisation types (Fig. 6).

2.2 Diversity indicators and metrics

We adapt the Shannon index (Gómez et al., 2024) to quantify diversity across five dimensions: field of study (FDI), language (LDI), gender (GDI), nationality (NDI) and affiliation type (ADI). For a given dimension, let S denote the number of classes and p_i the proportion of authors or articles belonging to class i. The standard Shannon index $H' = -\sum_{i=1}^{S} p_i \ln p_i$ captures richness and evenness but assumes that individuals belong to a single class. We normalise indices by dividing by $\ln S$ to facilitate comparison across dimensions. We compute indices annually for the 2010–2024 window and analyse temporal trends.

3 Dataset

The *DivinWD* dataset is available on Zenodo at `https://zenodo.org/records/18234750`.[5] The source code for the data pipeline is available at `https://github.com/DivinWD/dataset-resources` and it is released under the MIT license. Data and other content is released under the Creative Commons CC0 dedication (public domain). The datasheet for the dataset is available at `https://divinwd.dev/datasheet`.

4 Results

Figure 1 illustrates the exponential growth of Wikidata's scholarly corpus over three centuries, with a notable surge after 2010. However, the number of articles drops sharply in 2023 and 2024, reflecting the lag between publication and ingestion in Wikidata. Our diversity analysis covers publications from 2010 to 2024 and consists of 758,991 articles and 766,572 distinct authors.

4.1 Languages and Fields of Study

As shown in Figure 2, nearly 90% of the articles in our sample are written in English. Among non-English works the largest share is in Chinese, followed by Romance languages (French, Spanish, Italian). The low language diversity index (LDI ≈ 0.1, Fig. 7 (a)) reflects this dominance. Out of all articles, 1,120,505 could be assigned a field of study. Medicine is by far the most frequent discipline, followed by biology, environmental science, and physics; the humanities and social sciences are sparsely represented. The field-of-study diversity index (FDI, Fig. 7 (b)) remains moderate (around 0.45 when normalised) and shows little change over time, indicating persistent concentration in a handful of domains.

4.2 Gender, Nationality and Affiliation

Gender coverage in Wikidata is limited: roughly two-thirds of authors lack an explicit gender statement. Using data from the Genderize API, we find that male researchers outnumber female researchers by more than two to one, while non-binary classifications occur in less than 0.1% of cases, as shown in Figure 3. The gender diversity index (NDI, Fig. 7 (c)) increases slowly over the 2010–2024 period but remains low, mirroring structural imbalances in academia and incomplete data entry.

Nationality was inferred for 744,247 authors in total, 71,831 of whom had valid citizenship statements. As shown in Figure 4, Europe and North America dominate the authorship landscape, with Italy, Germany, and Poland being the most represented countries in this block (Fig. 4). Asian contributors, largely from China, constitute a smaller share, and authors from South America, Africa and Oceania are even less visible. The yearly trends for the top-10 countries are shown in Figure 5. The nationality diversity index (NDI, Fig. 7 (d)) remains relatively stable over time, reflecting enduring geographic disparities in the Wikidata corpus.

Author affiliations reveal that two ROR categories—*Education* and *Funder*—account for the majority of institutional types and follow similar growth trajectories, as shown in Figure 6. The affiliation diversity index (ADI, Fig. 7 (e)) fluctuates slightly across years but does not exhibit a strong trend, suggesting that the institutional mix of authors in Wikidata remains broadly constant.

5 Conclusions

This exploratory analysis shows that Wikidata can serve as a valuable open platform for computing diversity indicators of scientific publications, but it also exposes significant limitations. Although more than 45 million articles are present in Wikidata, only a subset has complete author and metadata information. Diversity indices computed from the 2010–2024 corpus highlight the strong dominance of the English language, medicine and hard sciences, male authors, Western nationalities and academic/funding institutions.

Ultimately, *DivinWD* demonstrates that open knowledge graphs offer a promising path towards transparent, reproducible and equitable evaluation of scholarly diversity, provided that the underlying data are curated with inclusivity in mind.

[4] Genderize (`https://genderize.io`, accessed 2026-03-03) is a commecial web API that predicts a person's likely gender based on their first name, using large-scale aggregated data and returning probabilities and sample counts.

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

[Gómez et al.2024] Emilia Gómez, Lorenzo Porcaro, Pedro Frau, and João Vinagre. 2024. Diversity in artificial intelligence conferences. Technical report, European Commission, Joint Research Centre, Luxembourg. JRC137550.

[Nielsen et al.2017] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. 2017. Scholia, scientometrics and Wikidata. In *European Semantic Web Conference*, pages 237–259. Springer.

[Rasberry et al.2019] Lane Rasberry, Egon L. Willighagen, Finn Aring;rup Nielsen, and Daniel Mietchen. 2019. Robustifying scholia: paving the way for knowledge discovery and research assessment through wikidata. *Research Ideas and Outcomes*, 5:e35820.

[Willighagen et al.2026] Egon L Willighagen, Daniel Mietchen, Peter Patel-Schneider, Konrad Linden, Johannes Kalmbach, Lars G Willighagen, Wolfgang Fahl, and Hannah Bast. 2026. Scholia 2026: Compliance with sparql 1.1.

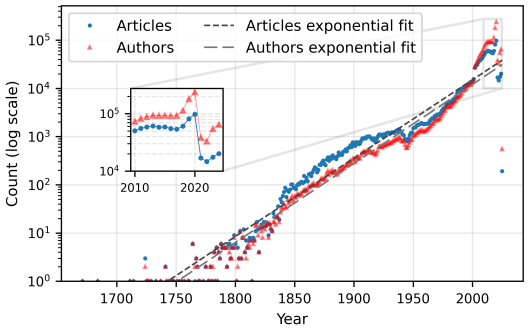

Figure 1: Annual article and author counts over time. Points show observed data, while dashed and solid lines represent exponential regression fits. Articles: $y = 4.7 \cdot 10^{-29} \cdot \exp(0.037 \cdot x)$, $R^2_{\text{articles}} = 0.94$, authors: $y = 1.9 \cdot 10^{-29} \cdot \exp(0.038 \cdot x)$, $R^2_{\text{authors}} = 0.95$.

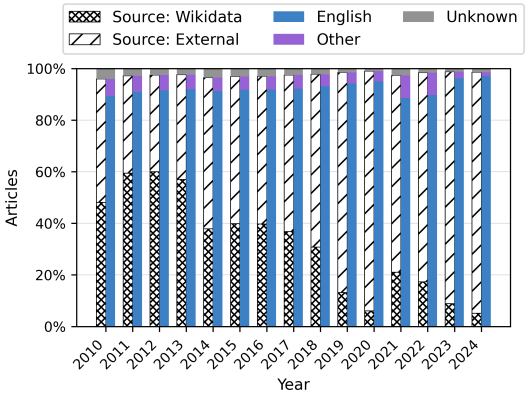

Figure 2: Distribution of articles' data source and language by year.

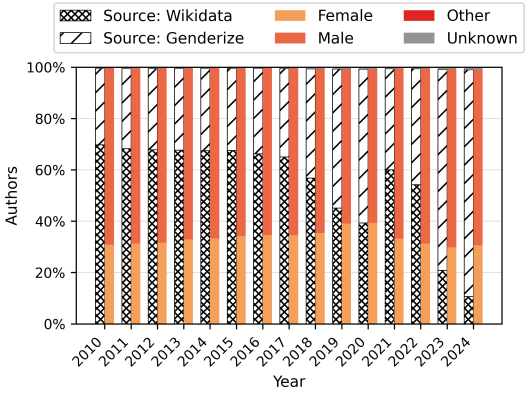

Figure 3: Distribution of author genders by year. The bars on the left show the source of the data.

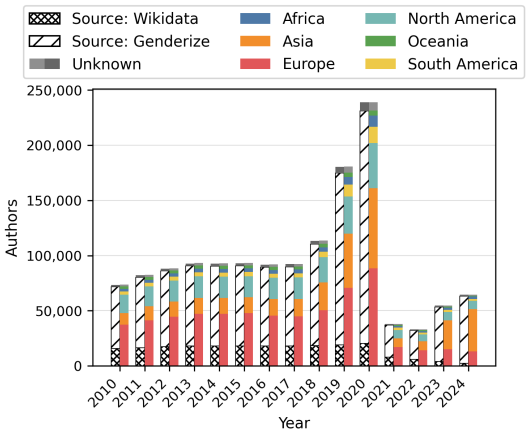

Figure 4: Distribution of authors by continent and by year. The bars on the left show the source of the data. Countries are aggregated into continents according to the Our World in Data mapping

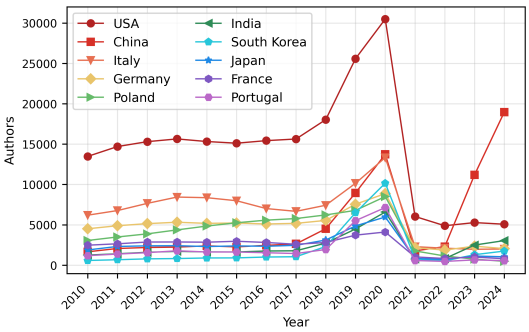

Figure 5: Top-10 countries by author citizenship in the period 2010-2024.

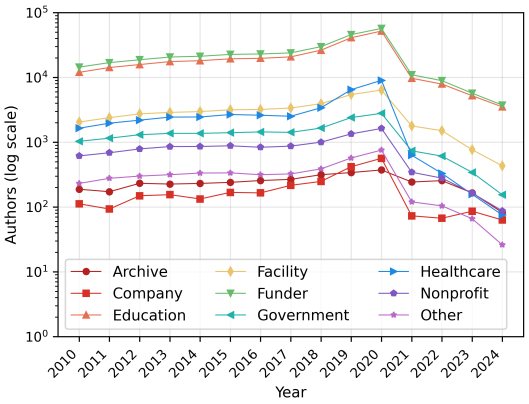

Figure 6: Number of authors by affiliation category between 2010 and 2024, shown on a logarithmic scale.

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.

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

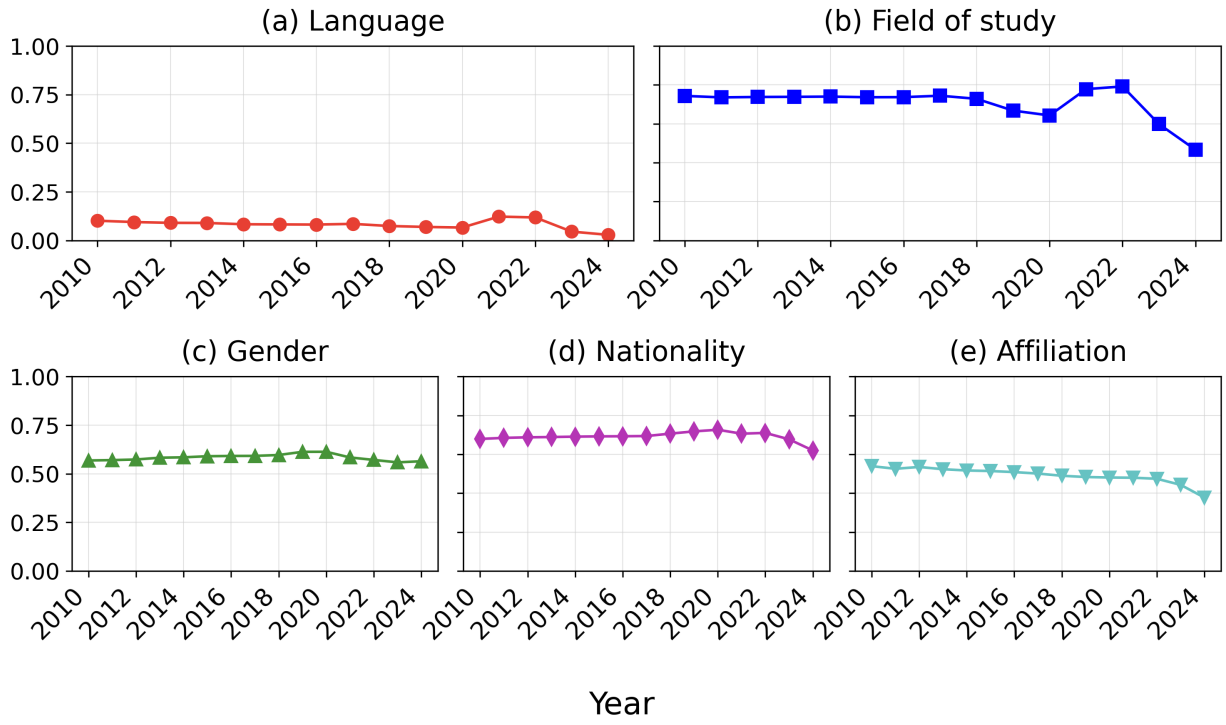

Figure 7: Shannon indices for language (a) and field of study (b) of articles, and for gender (c), country (d), and affiliation (e). The latter does not include type 'funder'. Each index is normalized in the range $[0, 1]$ by dividing H' by its maximum value $\ln |S|$. Specifically, the total number 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: $|S| = 8$.