# Project Aına: Understanding Data Visualization Research through Study Participants

Noëlle Rakotondravony *
Worcester Polytechnic Institute (WPI)

Jiarong Feng *
WPI

Yannick Metz †
ETH Zürich

Katy Williams ‡
Davidson College

Lane Harrison *
WPI

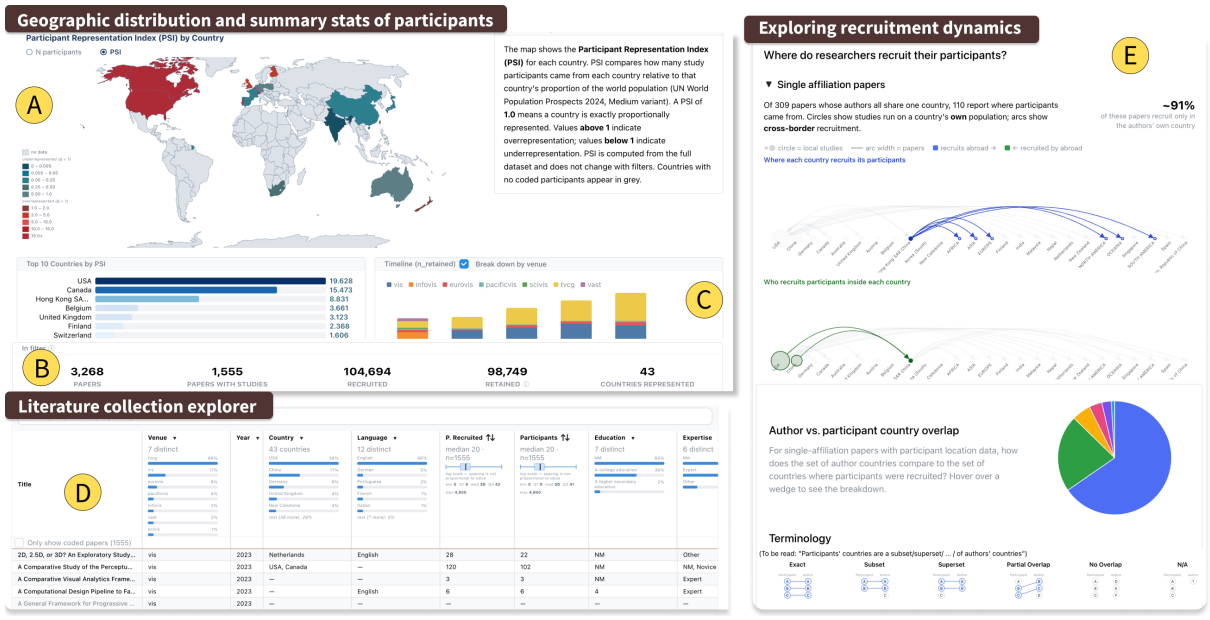



Figure 1: Main pages and sections of the Aına interface (aina.rakotondravony.com). Aına means "mirror" in Hindi, signifying that our participants are a reflection of who we choose to include in our research. (A) geographic distribution of study participants, (B) aggregate of codebook variables, (C) participants by venue over the years, (D) list of papers and their coded attributes, and (E) relationships between authors' affiliations and the populations they study.

## Abstract

Study participants play a critical role in the advancement and validation of visualization research. However, we currently lack a collective understanding of who our study participants are, how they shape research findings, and who may be left out. To address this gap, we present Aına *(eye-nah)*, a project that aims to contribute a dataset and an interactive interface for exploring the diversity of study participants in data visualization research published at VIS, EuroVis, PacificVis, and TVCG. The current data includes demographic and geographic details about participant samples and their involvement in studies, from 1555 visualization research papers. Our goal is to make this information accessible, understandable, and useful for the community from whom we seek feedback to better understand and develop the requirements towards this objective.

**Index Terms:** Study participants, meta-analysis, publication data

## 1 Introduction

In visualization research's long history of empirical evaluation, the methodologies, theories, and systems researchers develop are assessed through studies involving samples of the target users. Study participants thus greatly inform research outcomes, and are also a reflection of who we, as a community, ultimately serve. Understanding who is represented becomes necessary for contextualizing results, especially as visualization is adopted across domains and must reflect a variety of users [4] and researchers [2]. Yet obtaining an overview of human subjects' involvement remains challenging, as the relevant information is scattered across individual papers. A team designing a study may want to know whether similar populations were recruited before, or whether their target population fills a gap. Meta-researchers and those outside our community may want to assess how representative samples are: to what extent research reaches beyond students and participants from high-income countries, and what factors influence recruitment. In each case, the same question persists: *who has been studied, and who has not?* Current digital libraries do not support exploring it systematically.

We present project Aına, a work in progress that makes information about study participants accessible and easy to grasp, through two contributions: the Aına **dataset**, a structured record of participants reported in research published at VIS, EuroVis, PacificVis, and TVCG; and the Aına **interface**, a visual analytics tool for exploring it. While we focus on visualization studies, we believe the benefits of our work could transfer to other human-centered fields. We use the poster session to gather community feedback on both resources and on the use cases they should ultimately serve.

Our work aligns with efforts that assemble large collections of visualization publications for meta-analysis and research trend exploration. Such works focus largely on bibliographic metadata (*e.g.*, Vispubdata [3] and VisPubs.com [5]), while others character-

*e-mail: {ntrakotondravony, jfeng1, ltharrison}@wpi.edu
†e-mail: yannick.metz@inf.ethz.ch
‡e-mail: kawilliams@davidson.edu

ize authorship, collaboration, and gender trends [2]. Recent works also examined participants' expertise [1] and their involvement in visualization literacy tests [9]. These works highlight that studied populations are narrowly defined and limited in diversity, but each looks at the field through a specific lens, whereas human-subject involvement spans many areas and topics. The Aına dataset instead builds on human-centered fields that scrutinize participant diversity, most notably Linxen et al.'s seminal study of CHI publications [6]. Through both the dataset and the interface, we make participant information accessible and easy to grasp.

## 2 Method and Prototype

**Aına Dataset** The current dataset covers 1555 full papers published at VIS, EuroVis, PacificVis, and TVCG between 2020 and 2025 that report user studies or human participants outside the authorship list. It combines metadata from prior datasets (*e.g.*, [3, 5]) with information about the reported studies. Extending Linxen et al.'s methodology [6], we extract three categories per paper through systematic coding: **participant demographics**, **study information**, and **author information**.

Two coders first built a codebook by coding 120 randomly selected papers, adapting a scheme from Linxen *et al.* [6]. Following an inductive approach, they added variables relevant in prior literature, such as participants' domain expertise and the language of the research (*e.g.*, [1, 8]). After all authors reviewed and agreed on the scheme, four coders coded the papers in random order by publication year (2024 first, then 2023, etc.). The final codebook contains 12 codes; unreported variables were coded as "Not mentioned". To check consistency [7], each coder recoded a random 10% of articles coded by others; the average Cohen's $\kappa$ was 0.75 (substantial agreement with minimal subjectivity). Coders also met regularly to reconcile ambiguous cases and refine definitions—*e.g.*, splitting shared country counts evenly (50 participants from Canada and the USA $\rightarrow$ 25 each). The unit of analysis is the participant sample recruited from one country within one paper, so a paper spanning multiple studies or countries contributes multiple samples. As reporting varies, each analysis uses only papers reporting the relevant attribute (*e.g.*, participant country is available for 861 samples).

**Aına Interface** We designed the interface with three groups of users in mind, deriving their goals from anticipated usage scenarios: **Researchers** checking whether populations similar to their sample have already been studied, and where gaps remain (see Req 1, Req 2 below), **Meta-researchers** assessing how representative participation has been across the community, how it has evolved over time, and which factors correlate with recruitment ($\rightarrow$ Req 1, Req 3), **Practitioners from other fields** consulting the project as evidence of the populations on which existing findings rest and judging whether prior results transfer to their context ($\rightarrow$ Req 4).

We translate these scenarios into requirements (Req) for the current prototype: ✔ **Req 1. Provide corpus-level aggregation and overview**: Summarize overall participant representation, so that broad patterns are visible at a glance, ✔ **Req 2. Support targeted exploration**: Let users focus on a subset of papers of interest by filtering on attributes such as year, venue, and study characteristics, ✔ **Req 3. Surface recruitment dynamics**: Show relationships between who conducts research and who is studied, to help explain which populations become represented, ✔ **Req 4. Ensure data quality and transparency**: Make each data point traceable to its source paper, so that users can judge and trust what they see.

**The landing dashboard** delivers the corpus-level overview (Req 1) and targeted exploration through linked filters (Req 2), combining the participant map, summary statistics, per-venue timeline, and source-linked paper table (Fig. 1 A–D ; Req 4).

**The recruitment page** (Fig. 1 E) relates author affiliations to the populations they study (Req 3), helping reason about how collaborations and the location and modality (in-person vs. online) of studies and venues shape representation.

## 3 Preliminary Findings

In the current corpus of hand-coded papers starting in 2025 and back to 2020, we find that study participants are drawn from only 43 of the world's 195 countries. USA alone accounts for 73% of participants whose country we could identify, an over-representation of roughly 17 times relative to its share of world population, and a few countries—USA, Germany, the UK, and Canada—supply most recorded participants. Studies are overwhelmingly conducted in English (86.29% of samples with known language), the large majority of participants have at least a college education (81%), and about 92.67% of papers recruit participants locally, from the authors' own country.

## 4 Feedback We Seek from the Community

A natural limitation of our work is the field's lack of standardized participant reporting, and the prototype by itself does not prescribe the broad use cases the dataset and interface should ultimately serve. We seek the community's feedback on: (1) which use cases to support, as a top-down complement to our framing; (2) whether our four requirements and coded attributes are the right ones, and what is missing; and (3) whether the current encodings are clear and trustworthy. These conversations will shape our next steps: opening the dataset, supporting (semi-)automated coding, and grounding requirements in formative interviews with target users.


## Acknowledgments

We thank Daniel Keim and Priya Dhawka for their help in shaping the early ideas and values of the dataset. The dataset was created and coded manually by the team, and the interface prototype was developed with the help of Claude Code and Cursor.


## References


[1] A. Burns, C. Lee, R. Chawla, E. Peck, and N. Mahyar. Who Do We Mean When We Talk About Visualization Novices? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–16. Association for Computing Machinery, New York, NY, USA, Apr. 2023. doi: 10.1145/3544548.3581524 2

[2] H. Hao, Y. Cui, Z. Wang, and Y.-S. Kim. Thirty-Two Years of IEEE VIS: Authors, Fields of Study and Citations. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–10, 2022. doi: 10.1109/TVCG.2022.3209422 1, 2

[3] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, 2017. doi: 10.1109/TVCG.2016.2615308 1, 2

[4] A. Jena, M. Butler, T. Dwyer, K. Ellis, U. Engelke, R. Kirkham, K. Marriott, C. Paris, and V. Rajamanickam. The Next Billion Users of Visualization. *IEEE Computer Graphics and Applications*, 41(2):8–16, Mar. 2021. doi: 10.1109/MCG.2020.3044071 1

[5] D. Lange. Vispubs.com: A Visualization Publications Repository, May 2024. doi: 10.31219/osf.io/dg3p2 1, 2

[6] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, Yokohama Japan, May 2021. doi: 10.1145/3411764.3445488 2

[7] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, Nov. 2019. doi: 10.1145/3359174 2

[8] N. Rakotondravony, P. Dhawka, and M. Bancilhon. Beyond english: Centering multilingualism in data visualization. *arXiv preprint arXiv:2309.06659*, 2023. 2

[9] M. Solen. Scoping the future of visualization literacy: A review. *URL: osf. io/eypgm*, 8, 2022. 2