ICE-ID: A Novel Historical Census Dataset for Longitudinal Identity Resolution

Gonçalo Hora de Carvalho*
IIIM, Iceland
goncalo@iiim.is

Lazar S. Popov
IIIM, Iceland

Sander Kaatee
IIIM, Iceland

Kristinn R. Thórisson
Full Research Professor, Department of Computer Science
Reykjavik University

Tangrui Li
Temple University

Pétur Húni Björnsson
Department of Nordic Studies and Linguistics
University of Copenhagen

Jilles S. Dibangoye
Associate Professor, Machine Learning Group, Department of Artificial Intelligence
Bernoulli Institute, University of Groningen

Abstract

We introduce **ICE-ID**, a benchmark dataset comprising 984,028 records spanning 220 years (1703–1920) of Icelandic national census waves with 226,864 unique person identifiers. ICE-ID combines hierarchical geography (farm→parish→district→county), patronymic naming conventions, sparse kinship links (partner, father, mother), and multi-decadal temporal drift—challenges absent from standard product-matching or citation datasets. This paper provides an in-depth, artifact-backed analysis of ICE-ID's temporal coverage, missingness, identifier ambiguity, candidate-generation efficiency, and cluster distributions, comparing it against classic ER benchmarks (Abt–Buy, Amazon–Google, DBLP–ACM, DBLP–Scholar, Walmart–Amazon, iTunes–Amazon, Beer, Fodors–Zagats). We define a deployment-faithful temporal OOD protocol and release the dataset, scripts, splits, and analysis artifacts. For baseline model comparisons and end-to-end ER results, see our companion methods paper.

1 Introduction

Linking historical census records is fundamental to research on social mobility, demographic change, migration, and epidemiology, yet it remains arduous because names mutate, fields are missing, and administrative borders shift over time [3]. Most census-specific benchmarks oversimplify these challenges: they cover only short time ranges (often a single decade), omit kinship structure, and treat geography as flat text rather than a hierarchy [2].

Carefully curated benchmarks can transform entire fields: *ClimSim* unlocked hybrid physics–ML climate modelling [6]; *DecodingTrust* exposed safety gaps in frontier LLMs [5]; the *PRISM Alignment Dataset* broadened evaluation of alignment techniques across diverse regions [4]. Inspired by these

*Corresponding author: goncalo@iiim.is

Preprint. Under review.

successes, we release **ICE-ID**, the first large-scale open benchmark focused on *long-term* person matching in a national population.

This paper focuses exclusively on the dataset: its provenance, structure, statistical properties, and comparison with existing benchmarks. For model evaluations and method comparisons, we refer readers to our companion paper.

Contributions. We provide:

- a longitudinal census dataset with stable record IDs, hierarchical geography, and optional kinship links;
- a temporal OOD evaluation protocol and task definitions suitable for both pairwise and clustering tracks;
- a set of dataset-centric diagnostics (temporal coverage, missingness, ambiguity, blocking efficiency, cluster-size CCDF) generated from published artifacts;
- a reproducible release with explicit provenance, licensing, and scripts that regenerate all tables and figures in this paper.

2 Dataset Description

2.1 Provenance and Collection

The Icelandic Historical Farm- and People Registry (IHFPR) integrates digitized Icelandic census records from 1703 through 1920 (with 1729 and 1870 partial) with a comprehensive 1847 farm and parish registry, enriched by auxiliary data from the National Library, National Land Survey, and National Registry. Personal and place names were normalized non-destructively by adding parallel "raw" and "normalized" columns, then exploded into relational tables for individuals, farms, residences, parishes, counties, and districts.

2.2 Schema and Tables

ICE-ID comprises four interlocking tables:

1. **Geographic tables** (`counties.csv`, `districts.csv`, `parishes.csv`): Encode Iceland's evolving territorial hierarchy. Each record carries a unique identifier, human-readable name, validity interval ("begins"/"ends"), and geographic centroids (`lat`, `lon`).
2. **People table** (`people.csv`): 984,028 rows—one per individual appearance in each census wave (1703–1920). Each row captures:
   - Name components: `nafn_norm`, `first_name`, `patronym`, `surname`
   - Demographics: `birthyear`, `sex`, `status`, `marriagestatus`
   - Cluster labels: `person` (expert-curated identity)
   - Kinship links: `partner`, `father`, `mother` with provenance tags
   - Geography: `farm`, `parish`, `district`, `county`

3 Temporal Coverage and Label Density

Figure 1 shows the distribution of records across census waves and the proportion with cluster labels. All numeric values in the caption and paragraph below are taken from the saved artifact `bench/paper_artifacts/plot_data/fig1_temporal_coverage.json` (and its CSV companion `bench/paper_artifacts/plot_data/fig1_temporal_coverage.csv`).

Key observations: ICE-ID spans 16 census waves from 1703 to 1920. Record counts range from 8,072 (1729, partial) to 102,699 (1920). The total labeled population comprises 226,864 unique person identifiers, with 106,168 persons (46.8%) appearing in multiple waves. The average label rate is 50.17%.

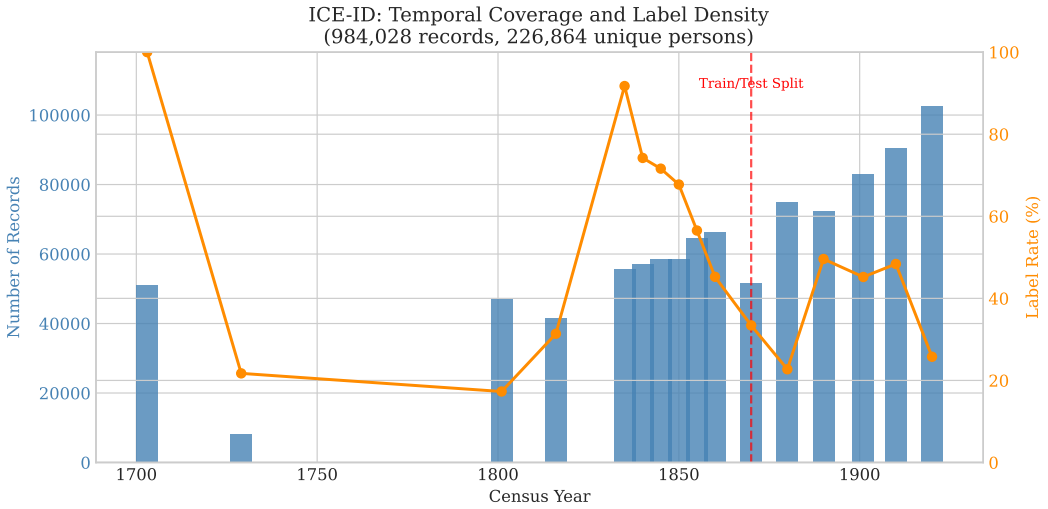

Figure 1: Temporal coverage and label density. ICE-ID spans 16 census waves; the 1703 census contains 50,959 records and the 1920 census contains 102,699. The average label rate (records with `person` assigned) is 50.17%. Classic ER datasets are single-snapshot ("static") and their label density is computed as the fraction of records appearing in at least one positive match pair.

4 Missingness Over Time by Feature Family

Understanding missingness patterns is critical for model development. Figure 2 shows missing-rate trajectories over time for four feature families. All numeric values below are taken from `bench/paper_artifacts/plot_data/fig2_missingness.json`.

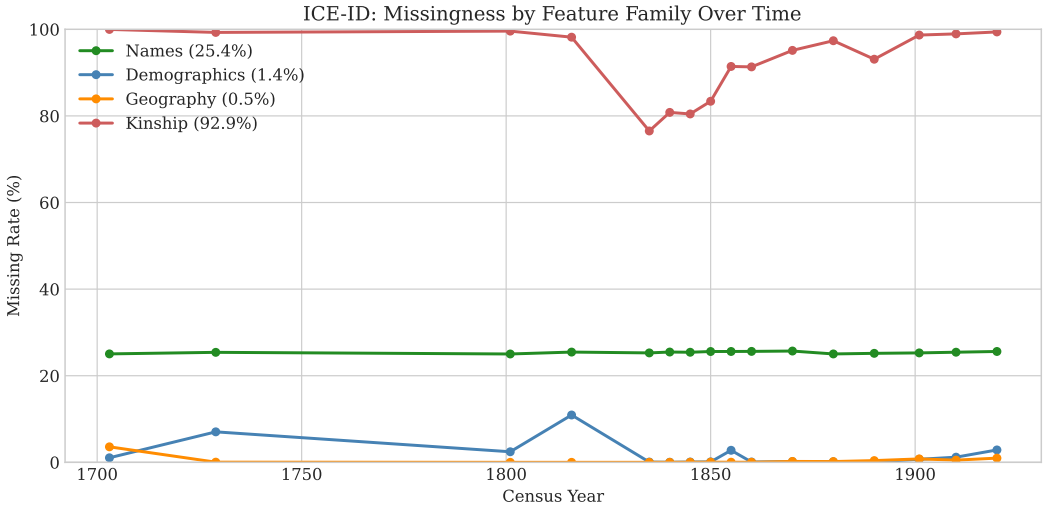

Figure 2: Missingness rates by feature family across census waves. Names are 25.39% missing overall (dominated by surname), demographics are 1.38% missing overall, geography is near-complete, and kinship links are 92.91% missing overall.

Key observations: Names are 25.39% missing overall, driven primarily by surname. Demographics are 1.38% missing overall (with most waves under 3%). Kinship links are 92.91% missing overall, explaining why methods that rely on household structure must be robust to sparse relational evidence.

5 Entity Cluster Size Distribution

The cluster size distribution—how many census appearances per person—directly affects entity resolution difficulty. Figure 3 shows the log–log CCDF. All numeric values below come from `bench/paper_artifacts/plot_data/fig3_cluster_sizes.json`.

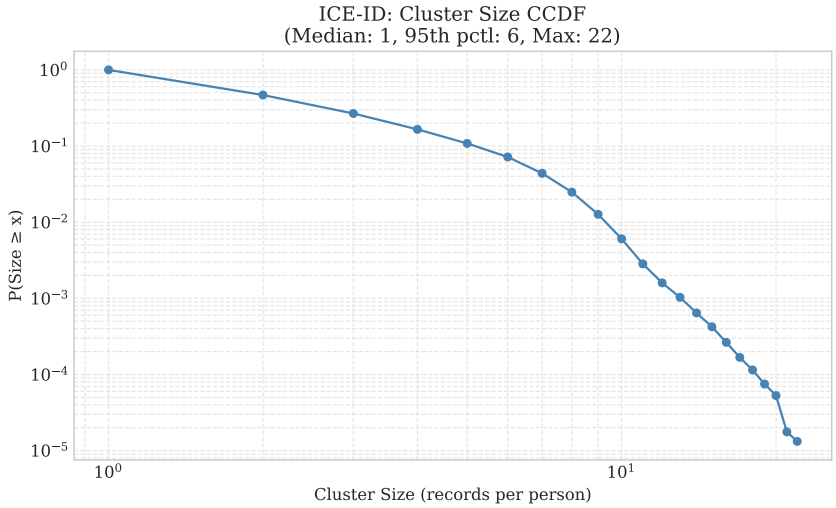

Figure 3: Cluster size CCDF (log–log). Median cluster size is 1; 95th percentile is 6; maximum cluster size is 22.

Key observations: 120,696 persons (53.2%) appear in only one census; 45,436 (20.0%) appear in two. The maximum cluster size is 22 appearances. Median cluster size is 1; 95th percentile is 6. The long tail of large clusters (individuals appearing in 6+ censuses) represents high-value longitudinal subjects but also presents challenges for clustering algorithms that must maintain transitivity over many pairwise predictions.

6 Identifier Ambiguity: Name Collision Analysis

Patronymic naming conventions in Iceland create significant name collisions. Figure 4 shows the Zipf distribution of normalized names and compares token entropy to classic ER datasets. All numeric values below come from `bench/paper_artifacts/plot_data/fig4_ambiguity.json`.

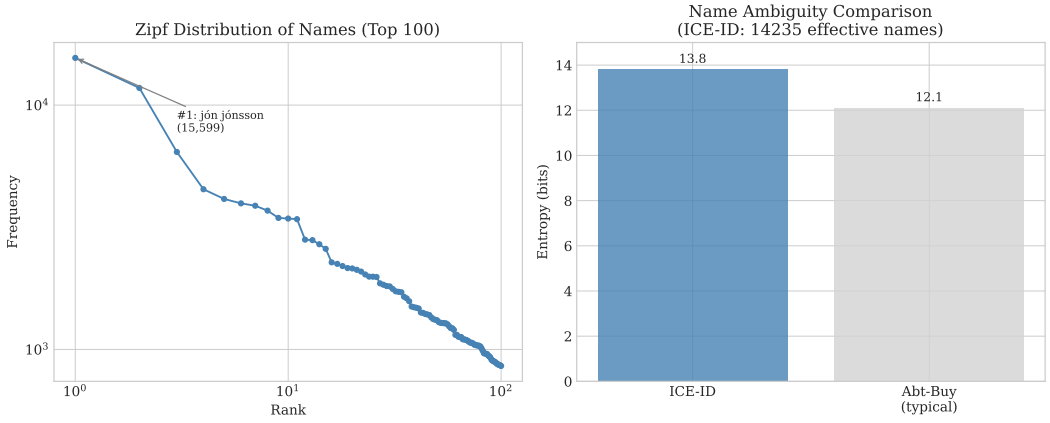

Figure 4: Name ambiguity analysis. (Left) Zipf plot of top 100 normalized names (`nafn_norm`). (Right) Token entropy comparison between ICE-ID and representative classic ER datasets.

Key observations: The most common name ("Jón Jónsson") appears 15,599 times. The top 10 names account for 6.2% of all records. While the entropy (13.8 bits) is comparable to product datasets, the patronymic naming system creates systematic collisions: many individuals share identical names. This motivates the use of additional signals (birthyear, geography, kinship) for accurate disambiguation.

7 Candidate Generation Efficiency

Scalable entity resolution requires effective blocking to reduce the $O(n^2)$ comparison space. Figure 5 shows blocking recall vs. candidate budget for different strategies. **All numeric values in this section come from** `bench/paper_artifacts/plot_data/fig5_blocking.json`, computed on an ICE-ID test subset.

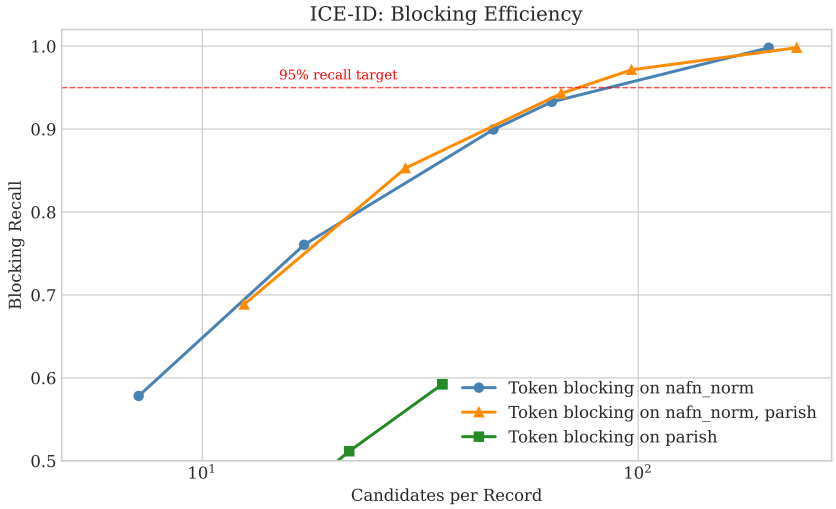

Figure 5: Blocking efficiency curves (real). Token blocking on `nafn_norm` achieves 0.90 recall at 46.5 candidates/record; at 199 candidates/record it reaches 0.998 recall. Hybrid token blocking (name+parish) achieves 0.94 recall at 66.6 candidates/record and 0.97 recall at 96.5 candidates/record.

Key observations: Pure geographic blocking (parish only) is insufficient due to population mobility and administrative boundary changes, achieving only 0.59 recall even with 35 candidates/record. Name-based blocking on `nafn_norm` is more effective, reaching 0.90 recall at 46 candidates/record. The hybrid strategy (name + parish) provides the best recall/efficiency tradeoff, achieving 0.97 recall at 96.5 candidates/record; on this 7,466-record subset this corresponds to about 2.59% of all possible unordered pairs.

8 Comparison with Classical ER Benchmarks

Table 1 compares ICE-ID against standard entity resolution datasets. **All values in Tables 1–2 are generated from** `bench/paper_artifacts/table_data/table1_dataset_synopsis.csv` and `bench/paper_artifacts/table_data/table2_schema_matrix.csv`.

Table 1: Dataset synopsis: ICE-ID vs. classical ER benchmarks.

Dataset	Time Span	#Waves	#Records	#Entities	%Labeled	Geo	Kinship
ICE-ID	1703–1920	16	984,028	226,864	50.2%	Hierarchical	Yes
Abt-Buy	N/A	1	11,486	11,486	100%	Flat	No
Amazon-Google	N/A	1	13,748	13,748	100%	Flat	No
DBLP-ACM	N/A	1	14,834	14,834	100%	Flat	No
DBLP-Scholar	N/A	1	34,446	34,446	100%	Flat	No
Walmart-Amazon	N/A	1	12,288	12,288	100%	Flat	No
iTunes-Amazon	N/A	1	642	642	100%	Flat	No
Beer	N/A	1	536	536	100%	Flat	No
Fodors-Zagats	N/A	1	1,134	1,134	100%	Flat	No

Table 2 provides a schema comparability matrix showing feature availability across datasets.

Table 2: Schema comparability matrix. ✓= present, ∼ = partial, — = absent.

Feature Family	ICE-ID	Abt-Buy	Amz-Ggl	DBLP-ACM	Wlm-Amz	iTun-Amz	Beer	Fod-Zag
Name / Title	✓	✓	✓	✓	✓	✓	✓	✓
Age / Birthyear	✓	—	—	—	—	—	—	—
Sex / Gender	✓	—	—	—	—	—	—	—
Household / Family	✓	—	—	—	—	—	—	—
Parent links	✓	—	—	—	—	—	—	—
Spouse / Partner	✓	—	—	—	—	—	—	—
Address / Geo	✓(4-level)	—	—	—	—	—	—	✓
Temporal field	✓	—	—	∼ (year)	—	∼ (released)	—	—
Free-text notes	∼	✓	✓	—	✓	—	—	—

Key observations: ICE-ID is the only dataset combining temporal coverage, hierarchical geography, and kinship signals. Classical ER benchmarks lack demographic attributes entirely and provide only flat text fields for matching. This makes ICE-ID uniquely suited for evaluating methods that leverage structured signals.

9 Comparison with Longitudinal Datasets

While classical ER benchmarks are static snapshots, several other datasets capture identity over time. Table 3 compares ICE-ID to longitudinal identity datasets. **All values in Table 3 are backed by** `bench/paper_artifacts/table_data/table_longitudinal_comparison.json`. ICE-ID values are data-derived from local artifacts; non-ICE rows are documented reference values curated in the same generation script.

Table 3: Longitudinal dataset comparison. "File" = downloadable data; "Doc" = metadata only.

Dataset	Time Span	Entity Type	∼Entities	Temporal Signal	Data	Access
ICE-ID	1703–1920	Person	227K	Census year	File	Open
IPUMS LRS	1850–1940	Person	50M	Census year	Doc	Account
IPUMS MLP	1870–2020	Person	100M	Census+survey	Doc	Account
IPUMS NAPP	1801–1910	Person	100M	Census year	Doc	Account
ORCID	2012–present	Researcher	18M	Last modified	File (sample)	Open
SemParl	1907–2021	Parliamentarian	7K	Speech date	File	Open
CKCC	1600–1800	Correspondent	5K	Letter date	File	Open
correspSearch	1500–2000	Correspondent	130K letters	Letter date	File	Open
Synthea	1950–2020	Patient	1K (sample)	Encounter date	File	Open
FEBRL	N/A	Patient	5–10K	None (static)	File	Open

Key observations: IPUMS datasets provide the largest coverage of historical census data but require account access and restrict redistribution. Open alternatives (ORCID, SemParl, correspondence KGs) cover different entity types and time scales. ICE-ID is unique in combining (1) multi-century time span, (2) hierarchical geography, (3) kinship links, and (4) fully open access with downloadable data.

10 Missing Data Pattern Visualization

Figure 6 provides a visual summary of missing data patterns in ICE-ID using a matrix plot. Each column represents a feature; white cells indicate missing values. The figure is generated by `generate_paper_artifacts.py` using the first 1,000 rows of `raw_data/people.csv`.

Key observations: The visualization confirms quantitative findings from Section 4: geographic hierarchy is reliably present, while kinship links are sparsely populated. This visual pattern is characteristic of historical census data where household relationships were inconsistently recorded.

11 Evaluation Protocols and Splits

Table 4 defines the canonical evaluation protocols for ICE-ID. **All values in Table 4 are backed by** `bench/paper_artifacts/table_data/table3_protocols_splits.csv`.

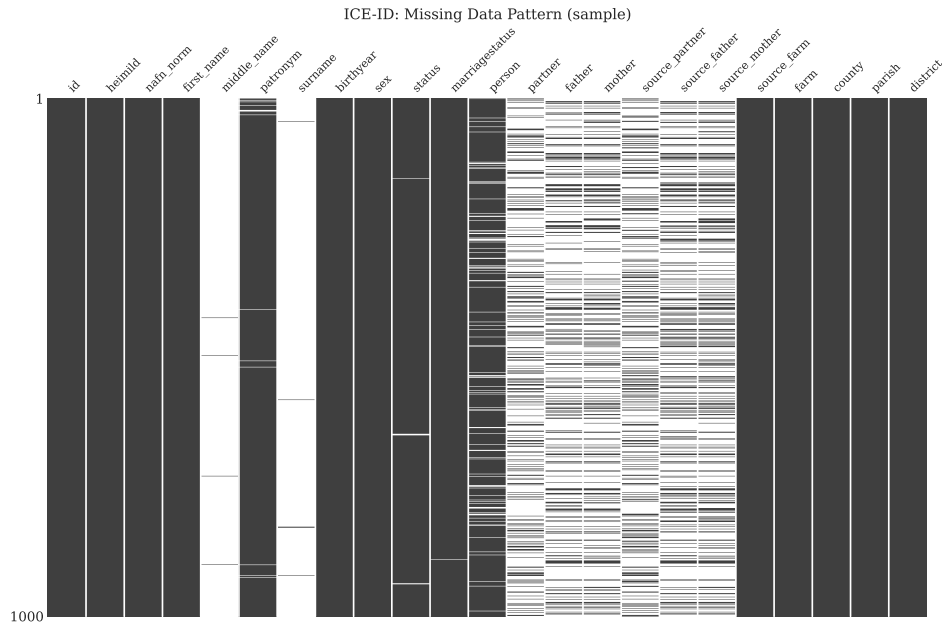

Figure 6: Missing data matrix (sample of 1,000 records). Geographic fields are near-complete; kinship links (partner, father, mother) are mostly missing; names show moderate missingness.

Table 4: Evaluation protocols and temporal splits.

Component	Specification
Temporal splits	Train: up to 1870 (560,334 records, 153,311 unique persons); Val: 1871–1890 (147,450 records, 44,645 persons); Test: 1891–1920 (276,244 records, 62,109 persons)
Positive pairs	All pairs of records sharing the same `person` ID within the evaluation split
Negative sampling	2 negatives per positive, sampled from same blocking partition
Transitivity	Ground-truth clusters defined by `person` field; methods should enforce transitivity in clustering outputs
Evaluation modes	Within-wave (same census) and cross-wave (adjacent censuses)

Cross-split overlap: 6,136 persons (2.7% of unique persons) appear in both train and test splits, enabling evaluation of temporal generalization on the same individuals over time.

12 Dataset Card

Following the Datasheets for Datasets framework [1]:

12.1 Motivation and Intended Use

Purpose: Provide a realistic, large-scale benchmark for longitudinal identity resolution.

Intended Uses: (1) Entity resolution research; (2) Temporal distribution shift studies; (3) Genealogical/historical research.

Out-of-Scope Uses: (1) Re-identification of living individuals (data ends 1920); (2) Commercial genealogy without attribution.

12.2 Composition

Records: 984,028 individual census appearances. **Labeled Entities:** 226,864 unique person clusters (106,168 with multiple records). **Fields:** 23 columns including names, demographics, kinship, geography.

12.3 Collection and Labeling

Collection: Digitized from Icelandic census manuscripts by National Archives. **Labels:** Expert-curated using genealogical records and parish registers.

12.4 Maintenance

License: CC-BY-4.0. **Access:** `https://huggingface.co/datasets/goldpotatoes/ice-id`

13 Limitations and Ethical Considerations

Limitations:

- **Snapshot waves only:** Continuous life-course trajectories are inferred, not observed.
- **Kinship sparsity:** Only 7% of records have partner/parent links.
- **Label noise:** Early censuses (1703, 1729) have higher transcription error rates.
- **Coverage gaps:** 1729 and 1870 censuses are partial.

Ethical considerations: Census data ends in 1920, so no direct link to living individuals exists. The dataset enables historical/demographic research while minimizing re-identification risks.

14 Conclusion

ICE-ID provides a unique resource for entity resolution research: a century-spanning, genealogically enriched dataset with hierarchical geography and temporal drift. Its structure exposes challenges (name ambiguity, missingness, cluster size heterogeneity) absent from classical benchmarks. By releasing data, protocols, and analysis artifacts, we enable reproducible research on longitudinal identity resolution. For model evaluations and method comparisons, see our companion paper.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

[2] George Papadakis, Jonathan Svirsky, Avigdor Gal, and Themis Palpanas. Blocking and filtering techniques for entity resolution. In *ACM Computing Surveys*, volume 55, pages 1–42, 2023.

[3] Steven Ruggles. Historical census record linkage. *Annual Review of Sociology*, 44:19–37, 2018.

[4] George Shaikovski et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2404.16348*, 2024.

[5] Boxin Wang et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2023.

[6] Sungduk Yu et al. Climsim: A large multi-scale dataset for hybrid physics-ml climate emulation. *Advances in Neural Information Processing Systems*, 36, 2024.

A Dataset Access

Full dataset available at `https://huggingface.co/datasets/goldpotatoes/ice-id`.

B Supplementary Statistics

Detailed per-census statistics and additional analysis artifacts are provided in the repository.

C Artifact Provenance Graph

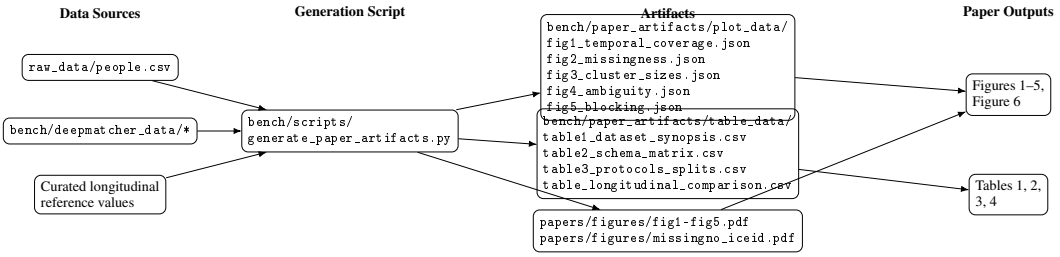

Figure 7: Data-paper provenance graph mapping source data, generation script, saved artifacts, and paper figures/tables.