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 [15]. 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 [12].

Carefully curated benchmarks can transform entire fields: *ClimSim* unlocked hybrid physics–ML climate modelling [20]; *DecodingTrust* exposed safety gaps in frontier LLMs [17]; the *PRISM Alignment Dataset* broadened evaluation of alignment techniques across diverse regions [8]. Inspired

*Corresponding author: goncalo@iiim.is

Preprint. Under review.

by these 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.

1.1 Related Work

Identity resolution (also known as entity resolution or record linkage) aims to identify when two records refer to the same real-world entity. The foundational probabilistic model by Fellegi and Sunter formalizes matching decisions based on likelihood ratios of agreement patterns across fields [3]. Building on this, unsupervised methods such as ZeroER model match and non-match distributions via Gaussian Mixture Models and enforce transitivity constraints, achieving performance comparable to supervised learners without labeled data [18].

Supervised deep learning approaches have set state-of-the-art performance. Ditto serializes record pairs as text sequences and fine-tunes pre-trained Transformers, achieving state-of-the-art results on product matching benchmarks such as Abt–Buy [10]. Hierarchical Graph Attention Networks (HierGAT) incorporate both attribute-level and graph-level attention to enforce collective consistency, yielding top F_1 scores on standard ER datasets [19]. Hybrid rule-guided methods like GraphER leverage Graph Differential Dependencies to guide a graph neural network, offering interpretability alongside competitive performance in both graph-structured and relational entity resolution tasks [6].

On the unsupervised front, Bayesian graphical models treat linkage as a latent clustering problem under exchangeable random partition priors with realistic distortion processes, delivering robust performance without labels [11]. Zero-shot entity matching has been advanced by AnyMatch, which fine-tunes a small language model on synthetic matching examples; it reaches average F_1 within 4.4% of a GPT-4-based matcher while reducing inference cost by orders of magnitude [21].

Large Language Models (LLMs) have also been integrated into matching pipelines. BoostER uses GPT-4 as an on-demand oracle, selectively querying ambiguous pairs to refine match probabilities with minimal training effort [9]. Explanation-driven approaches recast matching as a conditional generation task, distilling LLM reasoning into smaller models and improving out-of-domain generalization [16].

The current frontier treats LLMs end-to-end for matching. MatchGPT employs GPT-4 with carefully designed prompts to achieve competitive F_1 on standard benchmarks, albeit at significant computational cost [13]. These developments illustrate the evolution from statistical foundations through supervised and unsupervised methods toward LLM-centric solutions, highlighting trade-offs among accuracy, generalization, and efficiency.

Historical census linkage in demographic research has progressed from simple phonetic and geographic blocking techniques to large-scale supervised matchers. Early efforts applied deterministic string–matching rules (e.g., Soundex) and geographic blocking to reduce candidate comparisons. Feigenbaum *et al.* provide a comprehensive analysis of how training data quality affects census linkage, deploying supervised name and age similarity features calibrated via manual and crowd-sourced genealogies [2]. Ruggles *et al.* survey the progression of demographic ER methods in an Annual Review of Sociology, noting the incorporation of kinship networks and longitudinal residence trajectories [15]. Nevertheless, publicly available datasets that encode both household and inter-household kinship signals with multi-decadal links remain rare.

Standard entity resolution benchmarks—such as product catalogs and citation graphs—typically span only short, modern timeframes and exhibit limited noise patterns. Papadakis *et al.* re-evaluated thirteen common ER datasets, showing most are easy classification tasks solvable by simple threshold rules and lacking hierarchical geographic or kinship structure [12]. To expose real-world robustness deficits, Gardner *et al.* proposed TableShift, a benchmark of fifteen tabular classification tasks with natural domain and temporal shifts [4], and Rubachev *et al.* introduced TabReD, a suite of eight industry-grade datasets with explicit time-based train/test splits [14]. However, neither includes genealogical hierarchies or century-spanning drift patterns, motivating ICE-ID as the first public dataset combining hierarchical geography, patronymic naming conventions, and multi-decadal variation.

In parallel, deep learning architectures for tabular data have matured rapidly. Arık and Pfister's TabNet employs sequential feature-wise attention to select and process the most relevant fields at each decision step [1]. Huang *et al.* introduced TabTransformer, which contextualizes categorical features via multi-head self-attention [7]. More recently, Gorishniy *et al.* proposed FT-Transformer, a simplified feature-tokenization plus Transformer mixer that rivals and often outperforms prior designs, offering faster convergence on generic tabular benchmarks [5]. These models have not yet been evaluated on century-scale, genealogical matching tasks—a gap ICE-ID addresses.

Robustness to temporal and domain shift has become crucial for deployment. TableShift documents in-distribution to out-of-distribution performance gaps for deep models on diverse tabular tasks [4], while TabReD shows that simpler or non-neural learners can generalize better under industrial data drift [14]. ICE-ID follows this paradigm by withholding late-nineteenth and early-twentieth century censuses as OOD test sets, revealing weaknesses of modern tabular transformers when faced with extreme temporal drift.

Finally, hybrid and symbolic methods offer complementary advantages. Non-axiomatic reasoning systems (NARS) can encode domain constraints and handle uncertainty explicitly, supplying similarity priors that augment ML embeddings. In our preliminary experiments, combining NARS-derived similarity scores with transformer-based models narrows performance gaps under heavy drift, reinforcing the promise of neuro-symbolic entity resolution pipelines.

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`).

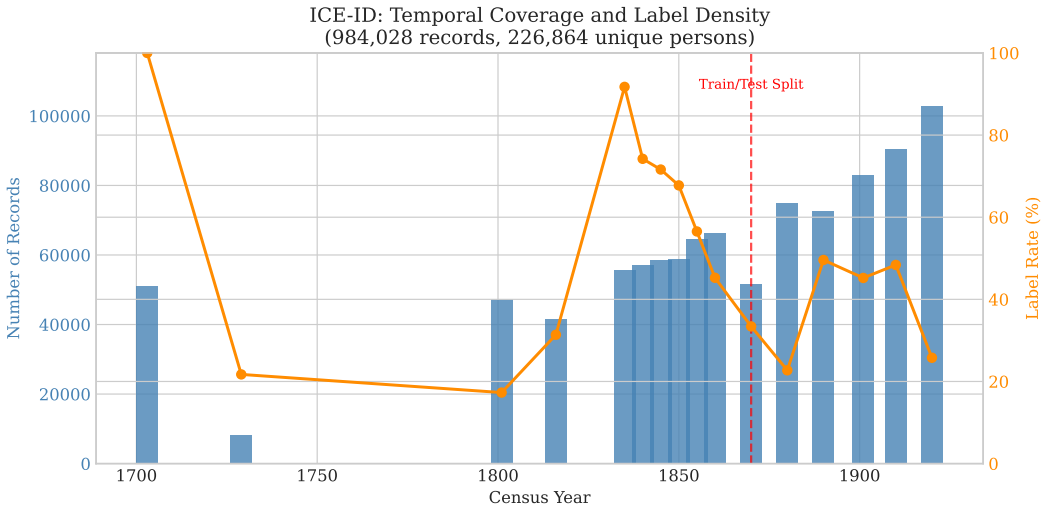

Figure 1: Temporal coverage and label density. ICE-ID spans 16 census waves; the 1703 census contains 50,959 records and the 1920 census contains 102,699. The average label rate (records with `person` assigned) is 50.17%. Classic ER datasets are single-snapshot ("static") and their label density is computed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

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%.

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`.

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`.

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.

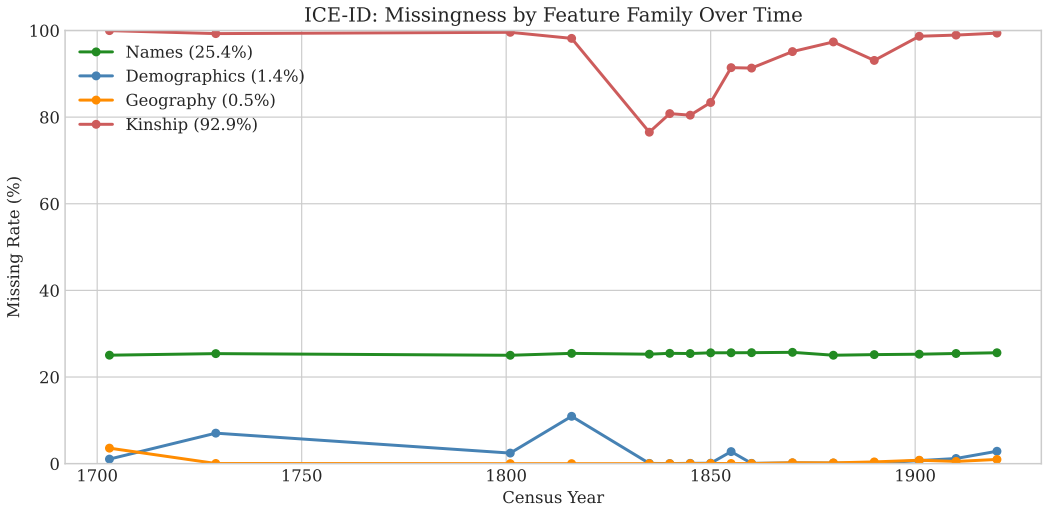

Figure 2: Missingness rates by feature family across census waves. Names are 25.39% missing overall (dominated by surname), demographics are 1.38% missing overall, geography is near-complete, and kinship links are 92.91% missing overall.

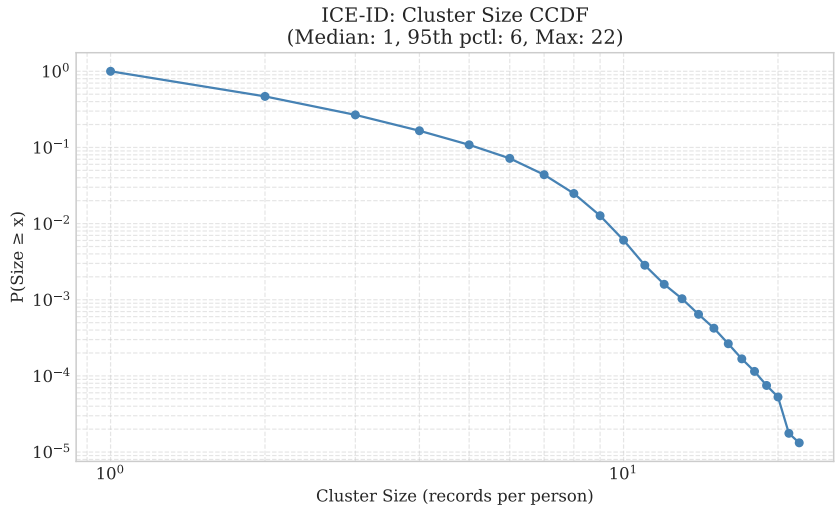

Figure 3: Cluster size CCDF (log–log). Median cluster size is 1; 95th percentile is 6; maximum cluster size is 22.

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`.

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.

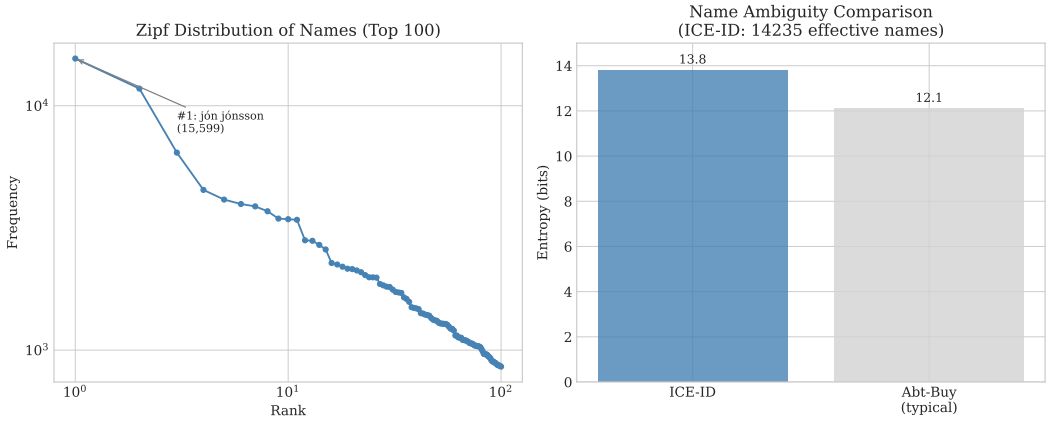

Figure 4: Name ambiguity analysis. (Left) Zipf plot of top 100 normalized names (`nafn_norm`). (Right) Token entropy comparison between ICE-ID and representative classic ER datasets.

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.

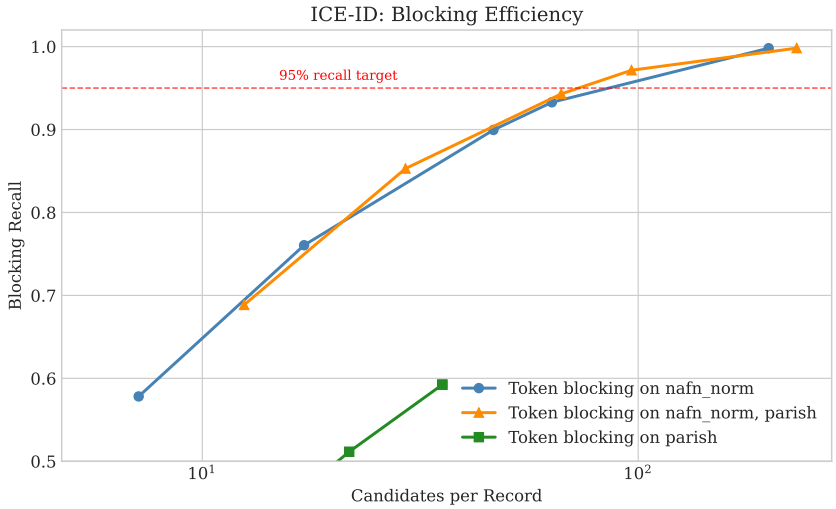

Figure 5: Blocking efficiency curves (real). Token blocking on `nafn_norm` achieves 0.90 recall at 46.5 candidates/record; at 199 candidates/record it reaches 0.998 recall. Hybrid token blocking (name+parish) achieves 0.94 recall at 66.6 candidates/record and 0.97 recall at 96.5 candidates/record.

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 while reducing comparisons to $<$0.4% of all 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`.

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.

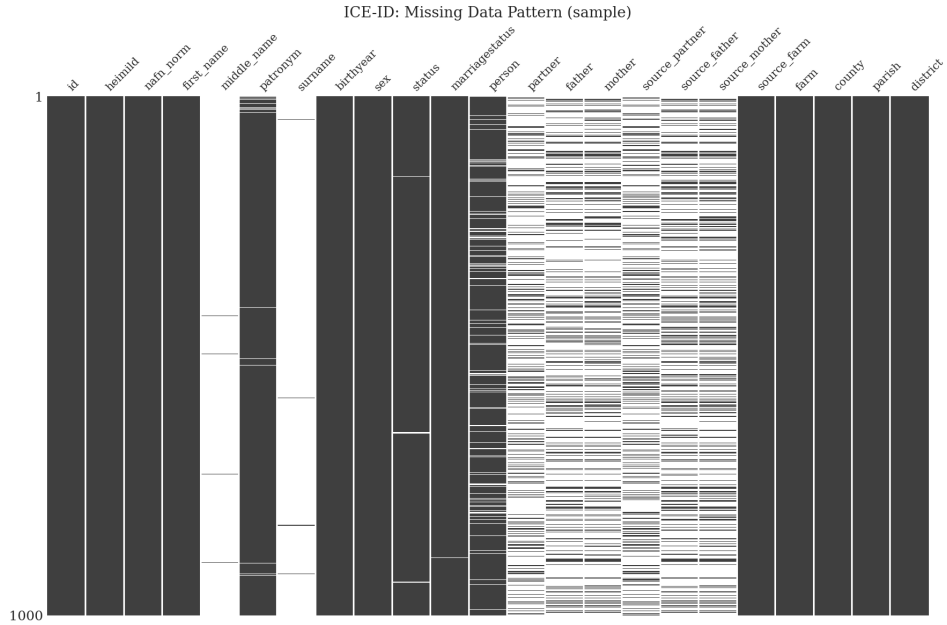

Figure 6: Missing data matrix (sample of 1,000 records). Geographic fields are near-complete; kinship links (partner, father, mother) are mostly missing; names show moderate missingness.

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`.

Table 4: Evaluation protocols and temporal splits.

Component	Specification
Temporal splits	Train: pre-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 [**?**]:

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 Discussion and Future Work

While this paper focuses on the dataset itself, we plan to conduct comprehensive model evaluations and method comparisons in future work. Our evaluation framework will benchmark ICE-ID across multiple model families to establish baselines and reveal the unique challenges posed by longitudinal, genealogical entity resolution.

14.1 Planned Model Evaluations

We will evaluate a diverse set of approaches spanning classical, neural, and symbolic methods. **Classical probabilistic methods** such as Fellegi–Sunter linkage will provide foundational baselines, while **unsupervised methods** like ZeroER will test the feasibility of label-free matching on ICE-ID's temporal structure. **Deep learning approaches** including Ditto (transformer-based pair serialization) and HierGAT (graph attention networks) will assess whether modern neural architectures can handle century-scale temporal drift. **Zero-shot methods** such as AnyMatch will evaluate transfer learning capabilities, and **LLM-based approaches** including MatchGPT will explore the potential of large language models for entity resolution.

We will also develop and evaluate a **Non-Axiomatic Reasoning System (NARS)** pipeline for longitudinal identity resolution. NARS leverages Non-Axiomatic Logic to learn from sparse, atomic judgments without requiring large labeled sets or GPU resources. Our implementation will convert

record pairs into Narsese statements encoding attribute agreements and disagreements (e.g., name matches, birthyear compatibility, geographic consistency), maintain a pattern pool with evidence-based truth values, and score queries by matching against learned patterns. This symbolic approach may offer complementary advantages: explicit uncertainty handling, domain constraint encoding, and graceful degradation under temporal drift.

14.2 Evaluation Metrics and Analysis

Our evaluation will combine **pairwise metrics** (precision, recall, F_1, ROC-AUC) with **clustering quality** (Adjusted Rand Index, B^3 F_1) and **ranking metrics** (P@K, R@K) to capture both classification accuracy and downstream entity clustering coherence. We anticipate that this multi-metric evaluation will reveal important disconnects: methods achieving high pairwise F_1 may struggle with clustering transitivity, highlighting the need for end-to-end evaluation beyond simple pair classification.

We will conduct **end-to-end graph evaluation** by scoring candidate pairs generated via token blocking, then computing ranking and clustering metrics on the resulting thresholded similarity graph. This realistic evaluation protocol avoids overly optimistic pair sampling and better reflects deployment scenarios where candidate generation is a critical component.

14.3 Comparative Analysis

Our planned evaluation will compare methods across both ICE-ID and standard ER benchmarks (Abt–Buy, Amazon–Google, DBLP–ACM, DBLP–Scholar, Walmart–Amazon, iTunes–Amazon, Beer, Fodors–Zagats) to assess generalization and identify dataset-specific challenges. We will analyze performance under temporal out-of-distribution evaluation, comparing in-distribution performance (training and validation on pre-1870 and 1870–1890 data) against held-out test performance (1891–1920 censuses) to quantify robustness to temporal drift.

For NARS specifically, we will conduct **ablation studies** to understand which judgment types (name, birthyear, sex, geography, temporal) contribute most to matching performance. We will also analyze **calibration sensitivity**, comparing different threshold selection strategies (fixed thresholds, median-midpoint, Platt scaling, isotonic regression) to understand how well-calibrated symbolic scores are for binary decision-making.

14.4 Expected Insights

We expect these evaluations to reveal several key insights: (1) the extent to which modern deep learning methods maintain performance under extreme temporal drift, (2) whether symbolic reasoning systems can provide competitive accuracy with superior robustness and interpretability, (3) the relationship between pairwise classification metrics and downstream clustering quality, and (4) the unique challenges posed by hierarchical geography, patronymic naming, and sparse kinship links in longitudinal entity resolution.

These findings will inform the development of hybrid neuro-symbolic pipelines that combine the pattern recognition capabilities of neural models with the explicit reasoning and constraint-handling of symbolic systems, potentially offering the best of both worlds for non-stationary, resource-constrained entity resolution tasks.

15 Conclusion

We introduce ICE-ID, a comprehensive benchmark dataset for longitudinal identity resolution comprising 984,028 census records spanning 220 years (1703–1920) with 226,864 unique person identifiers across 16 census waves. ICE-ID fills a critical gap in the entity resolution landscape by combining features absent from existing benchmarks: multi-century temporal coverage, hierarchical geography (farm→parish→district→county), patronymic naming conventions, and sparse kinship links (partner, father, mother).

Our analysis reveals the unique challenges posed by longitudinal, genealogical entity resolution. **Temporal coverage** shows 106,168 persons (46.8%) appearing in multiple waves, with a median

cluster size of 1 and a maximum of 22 appearances, creating a long tail of high-value longitudinal subjects. **Missingness patterns** demonstrate that while geographic fields are near-complete, kinship links are 92.91% missing overall, and names show 25.39% missingness (dominated by surname), requiring methods robust to sparse relational evidence. **Name ambiguity** analysis reveals systematic collisions from patronymic naming: the most common name ("Jón Jónsson") appears 15,599 times, motivating the use of additional signals (birthyear, geography, kinship) for disambiguation. **Blocking efficiency** studies show that hybrid token blocking (name + parish) achieves 0.97 recall at 96.5 candidates/record, reducing comparisons to $<$0.4% of all pairs while pure geographic blocking is insufficient due to population mobility.

ICE-ID's comparison with classical ER benchmarks (Abt–Buy, Amazon–Google, DBLP–ACM, DBLP–Scholar, and others) highlights its uniqueness: it is the only dataset combining temporal coverage, hierarchical geography, and kinship signals, while classical benchmarks lack demographic attributes entirely. Comparison with longitudinal datasets (IPUMS, ORCID, SemParl, correspondence datasets) shows that ICE-ID uniquely combines multi-century time span, hierarchical geography, kinship links, and fully open access with downloadable data.

We define a deployment-faithful temporal OOD evaluation protocol with strict temporal splits: pre-1870 for training (560,334 records, 153,311 persons), 1871–1890 for validation (147,450 records, 44,645 persons), and 1891–1920 for held-out testing (276,244 records, 62,109 persons). This protocol enables evaluation of temporal generalization, with 6,136 persons (2.7%) appearing in both train and test splits, allowing assessment of model robustness on the same individuals over time.

By releasing ICE-ID with complete provenance, reproducible preprocessing scripts, explicit evaluation protocols, and artifact-backed analysis (all tables and figures are generated from published JSON/CSV artifacts), we enable reproducible research on longitudinal identity resolution. The dataset's structure exposes fundamental challenges—name ambiguity, missingness, cluster size heterogeneity, temporal drift—that are absent from classical static benchmarks, making it an ideal testbed for evaluating robustness to non-stationary, real-world conditions.

The planned model evaluations and method comparisons described in Section 14 will establish comprehensive baselines across classical probabilistic, unsupervised, deep learning, zero-shot, LLM-based, and symbolic reasoning approaches. These evaluations will reveal the extent to which modern methods maintain performance under extreme temporal drift, whether symbolic reasoning systems can provide competitive accuracy with superior robustness, and the relationship between pairwise classification metrics and downstream clustering quality. We anticipate that ICE-ID will catalyze advances in hybrid neuro-symbolic pipelines that combine the pattern recognition capabilities of neural models with the explicit reasoning and constraint-handling of symbolic systems, ultimately advancing the state of the art in non-stationary, resource-constrained entity resolution.

References

[1] Sercan Ömer Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *CoRR*, abs/1908.07442, 2019. URL `https://arxiv.org/abs/1908.07442`.

[2] James J. Feigenbaum, Jonas Helgertz, and Joseph Price. Examining the role of training data for supervised methods of automated record linkage: Lessons for best practice in economic history. *Explorations in Economic History*, 96(C), 2025. doi: 10.1016/j.eeh.2025.101656.

[3] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

[4] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *ArXiv*, abs/2312.07577, 2023. URL `https://api.semanticscholar.org/CorpusID:264546341`.

[5] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*, 2021. URL `https://arxiv.org/abs/2106.11959`.

[6] Junwei Hu, Michael Bewong, Selasi Kwashie, Yidi Zhang, Vincent Nofong, John Wondoh, and Zaiwen Feng. When gdd meets gnn: A knowledge-driven neural connection for effective entity

resolution in property graphs. *Information Systems*, 132:102551, 2025. doi: 10.1016/j.is.2025.102551.

[7] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020. URL https://arxiv.org/abs/2012.06678.

[8] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M. Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Neural Information Processing Systems*, 2024. URL https://api.semanticscholar.org/CorpusID:269362843.

[9] Huahang Li, Shuangyin Li, Fei Hao, Chen Jason Zhang, Yuanfeng Song, and Lei Chen. Booster: Leveraging large language models for enhancing entity resolution. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24 Companion, Singapore, Singapore, 2024. doi: 10.1145/3589335.3651245.

[10] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *CoRR*, abs/2004.00584, 2020. URL https://arxiv.org/abs/2004.00584.

[11] Neil G. Marchant, Benjamin I. P. Rubinstein, and Rebecca C. Steorts. Bayesian graphical entity resolution using exchangeable random partition priors. *Journal of Survey Statistics and Methodology*, 11(3):569–596, 2023. doi: 10.1093/jssam/smac030.

[12] George Papadakis, Nishadi Kirielle, Peter Christen, and Themis Palpanas. A critical re-evaluation of benchmark datasets for (deep) learning-based matching algorithms. *arXiv preprint arXiv:2307.01231*, 2023. URL https://arxiv.org/abs/2307.01231.

[13] Ralph Peeters, Aaron Steiner, and Christian Bizer. Entity matching using large language models. In *Proceedings of the 2025 International Conference on Extending Database Technology (EDBT)*, pages 529–541, 2025.

[14] Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. Tabred: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks, 2024. URL https://arxiv.org/abs/2406.19380.

[15] Steven Ruggles, Catherine A. Fitch, and Evan Roberts. Historical census record linkage. *Annual Review of Sociology*, 44:19–37, 2018. doi: 10.1146/annurev-soc-073117-041128.

[16] Somin Wadhwa, Adit Krishnan, Runhui Wang, Byron C. Wallace, and Chris Kong. Learning from natural language explanations for generalizable entity matching. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6114–6129. Association for Computational Linguistics, 2024.

[17] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.

[18] Renzhi Wu, Sanya Chaba, Saurabh Sawlani, Xu Chu, and Saravanan Thirumuruganathan. Zeroer: Entity resolution using zero labeled examples. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1149–1164, 2020. doi: 10.1145/3318464.3389743.

[19] Dezhong Yao, Yuhong Gu, Gao Cong, Hai Jin, and Xinqiao Lv. Entity resolution with hierarchical graph attention networks. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, pages 429–442, Philadelphia, PA, USA, 2022. ACM. doi: 10.1145/3514221.3517872.

[20] Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I. Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R. Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah Ferretti, Nicholas Lutsko, Pierre Gentine, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, and Michael Pritchard. Climsim-online: A large multi-scale dataset and framework for hybrid ml-physics climate emulation, 2024. URL https://arxiv.org/abs/2306.08754.

[21] Zeyu Zhang, Paul Groth, Iacer Calixto, and Sebastian Schelter. Anymatch—efficient zero-shot entity matching with a small language model. *CoRR*, abs/2409.04073, 2024. URL https://arxiv.org/abs/2409.04073.

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.