Rillor

EXPLAINER

How entity resolution matches records across sources

How entity resolution decides which records describe the same person or company: blocking, the Fellegi-Sunter rule, clustering, registry identifiers, language models and measuring the result.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Entity resolution decides which records describe the same person, company or product when sources share no reliable key. Exact joins miss many true matches: more than 25% in one census application, and Winkler estimates that in some business lists possibly 40% to 60% of true pairs differ significantly on both name and address [1].
- Blocking makes the comparison count tractable and caps recall. Ten passes kept 99.5% of true census matches in about one trillion of roughly 100 quadrillion possible pairs [1].
- The Fellegi-Sunter model scores each pair from field agreements and holds pairs between two thresholds for clerical review [1, 3]. Clustering then turns links into entities, and the method and threshold change the result [9, 32].
- The Global LEI Index held 3,487,659 records on 7 November 2026, including 6,518 marked as duplicate registrations, and the LEI does not cover people acting privately [20, 22].
- The strongest language model tested came within 4.49 F1 points of the best fine-tuned matcher, or beat it, on six public benchmarks without task-specific training [10]. Most established benchmarks are easy [12], and a 2026 study found language models' explanations of their match decisions often unstable and weakly faithful [14].
- Common measures can rank the same results in opposite order [7]. A result is judged by its truth set, threshold and more than one measure, with each decision kept with its score and pass [8, 16].

Background

Record linkage, in the words of a Census Bureau overview, "is the means of combining information from a variety of computerized files" [1]. Its methods compare names, addresses and other fields across pairs of records to find those that belong to the same entity, which "might be a business, a person, or some other type of unit that is listed" [1]. The same report lists other names for the field: object identification, data cleaning, approximate joins, fuzzy matching and entity resolution [1].

Binette and Steorts, in a 2022 review in Science Advances, draw the distinctions this report uses [5]:

Term	Definition in the source
Deduplication	Removing duplicate records within one database
Record linkage	Merging databases and removing duplicates across them. Bipartite linkage removes duplicates across databases but not within them
Entity resolution	Merging databases and removing duplicates across and within them at the same time
Canonicalization	Constructing one representative record per entity, resolving conflicting information

Usage in practice is loose. The Census Bureau's 2024 software comparison says entity resolution is "also referred to as record linkage or deduplication" [3], and language-model studies call the pairwise decision entity matching [10]. The pipeline is usually drawn in four stages: attribute alignment, blocking, record linkage and canonicalization [5]. Standardization sits at the front, for example replacing the variants of "Incorporated" with "Inc." [1].

Statistical agencies depend on the method. The Census Bureau used entity resolution in the 2020 Census to deduplicate responses and to link people listed in the Post-Enumeration Survey to people enumerated in the census [3]. Its research page, last revised 29 October 2026, lists among potential applications the linkage of administrative and commercial person-level lists for Census 2030 and unduplication for frame construction in the demographic and economic areas [2].

The model behind the Census Bureau's linkage software and Splink comes from Fellegi and Sunter (1969) [1, 3, 27].

Method

What was examined. Each stage of the pipeline: comparison on exact keys, blocking, pairwise scoring, clerical review and clustering. Two further topics decide how far a result can be trusted: what registry identifiers cover, and how results are measured. Language-model matchers were examined as a scoring method.

Sources. Thirty-five sources, each opened during this review: U.S. Census Bureau research [1, 2, 3, 4], peer-reviewed surveys and evaluation studies [5, 6, 7, 9], a practitioner's guide to evaluation posted on arXiv [8], matching studies published at EDBT, ICDE and PVLDB [10, 12, 13] plus three preprints [11, 14, 15], guidance on reporting linkage error [16, 17], GLEIF, LEI Regulatory Oversight Committee and SEC documentation and data [18, 19, 20, 21, 22, 23, 24, 25, 26], and the documentation of Splink, an open-source linkage package from the UK Ministry of Justice [27, 28, 29, 30, 31, 32, 33, 34, 35].

How. Each claim was checked against the passage in the opened source, and figures are given as stated, with dates. LEI counts were queried from the GLEIF API on 7 November 2026 against the golden copy published at 00:00 UTC that day [20]. The 1969 Fellegi and Sunter paper could not be opened from its publisher's site, so the model is described through sources that summarize it [1, 3, 29]. Example records are the illustrations printed in the sources. This report uses no Rillor data.

Findings

1. Exact joins miss many true matches

Typing errors alone defeat exact comparison. Winkler reports that in a major census application for measuring undercount, "more than 25% of matches would not have been found via exact character-by-character matching" [1]. In the St Louis region, 25% of first names and 15% of last names did not agree exactly among true matches [1]. Small error rates compound across fields: a 10% error rate on first name and 5% on last name and each birth-date field "could miss 25% of matches" in an exact join across quarters [1].

Businesses are harder than people, because of variant names and addresses [1]. Winkler's Table 2 shows one gasoline station listed under the name of its location, under its owner's name and home address, and as "J A S, Inc c/o Robert Jones, CPA" at the accountant's address [1]. In a number of business-list situations, "many (possibly 40-60%) of truly matching pairs have both names and address that differ significantly" [1].

Names change, and punctuation and legal forms differ by source. The SEC record for CIK 0001326801 reads "Meta Platforms, Inc." and lists the former name "Facebook Inc", without the period, from 6 May 2005 to 27 October 2021 [25, 26]. The ISO 20275 code list, in its February 2026 version, holds more than 3,600 entity legal forms across more than 200 jurisdictions, each only in its local language [24]. Fields go missing too: in one company-matching case, 60% of the records in an employer table had no zip code [13].

Rules alone do not close the gap. The Splink documentation shows three records for one person, written with a nickname, a full first name and a quoted nickname, where two deterministic rules find none of the matches. It calls deterministic linkage cheap and capable of high precision, and prone to low recall [31]. GUILD cites evidence that probabilistic linkage gives less biased results than deterministic linkage alone [16]. It also warns that cleaning, such as removing spaces from postcodes, can create false links as well as true ones, so standardization rules "should, therefore, be reported in detail" [16].

2. Blocking cuts comparisons and caps recall

Candidate pairs grow with the square of the input. Linking two files means comparing the size of one times the size of the other, and deduplicating a file means comparing every record with every other [3, 6]. The Splink documentation puts one million records at about 500 billion comparisons [28]. The 1,167,705 records in the Census Bureau's 2024 test give 681,766,899,660 possible pairs, among which the truth set holds 3,941 true links (arithmetic from [3]). The PVS paper calls all-pairs comparison "computationally untenable" for incoming files that can exceed 500 million records [4].

Blocking compares only pairs that agree on some characteristic. Newcombe introduced it with keys such as surname or date of birth, and Hernandez and Stolfo added multiple passes with different sort orders and a sliding window [1]. The blocking survey calls it "an inherently approximate solution" that trades slightly lower effectiveness for much higher efficiency [6]. Standard blocking "is very sensitive to noise", because the

slightest difference in two duplicates' keys puts them in different blocks [6]. Binette and Steorts add that traditional blocking on fields such as gender and year of birth "amounts to a deterministic a priori judgment that these fields are error-free" [5].

Practice answers with several independent passes. The PVS uses a strict first pass with a low chance of error, such as an exact address match, then looser passes, and records linked in one pass leave the later ones [4]. The Census Bureau's 2024 test ran BigMatch and Splink on the four-pass scheme of the 2020 Post-Enumeration Survey: phone number, then a census unit identifier with name initials, then initials with month and day of birth, then geography, name prefixes, sex and age group [3]. The third tool, fastLink, could not handle that volume of blocking and ran on a looser scheme with fewer, bigger blocks [3]. The paper says the survey's scheme scales to a nationwide deduplication task for BigMatch "while reducing the amount of falsely discovered links to produce a manageable clerical review workload" [3]. Splink recommends a longer list of strict rules over a short list of loose ones, so a typo in the name is caught when the postcode matches, and the reverse [28].

The recall ceiling can be high and is never complete. In Winkler's study of a 300-million-record census, ten passes reduced about 100 quadrillion pairs to about one trillion that held 99.5% of the true matches. The report elsewhere states the setup as 11 blocking criteria [1]. The missed 0.5% included children listed under two last names, with date of birth missing in one listing and street address in the other, whose records shared no three-character sequences. Winkler judged such pairs unlikely to be found by any computerized procedure using only the census files [1].

Newer blockers use similarity search. In the Ditto employer case, zip-code blocking was joined with the 20 most similar records by TF-IDF similarity of name and address, giving about 10 million candidates, and embedding-based blocking made the matching process 3.8 times faster [13].

Blocking has its own measures [6]:

Measure	What it measures
Pair Completeness	Share of true duplicates that survive blocking (recall of the blocks)
Pairs Quality	Share of compared pairs that are true duplicates (precision of the blocks)
Reduction Ratio	One minus the ratio of compared pairs to all possible pairs

By those definitions, Winkler's census example has a reduction ratio of about 0.99999 with pair completeness of 0.995 [1, 6].

3. The Fellegi-Sunter model scores each pair

The Census Bureau has used the Fellegi-Sunter framework for large-scale linkage and deduplication since the 1980s [3]. The model classifies pairs into true matches and true non-matches. For the observed pattern of agreement between two records, it takes the ratio of its probability among matches to its probability among non-matches. That ratio, or an increasing function of it such as its natural log, is the matching weight [1].

The Splink documentation names the parts. The m probability, the chance of an observation given a match, is "largely a measure of data quality". The u probability, the chance of the same observation given a non-match, is "a measure of coincidence". A prior, lambda, is the probability that any two records match [29]. Because fields are assumed independent, Bayes factors multiply and log weights add [29]. The 2024 Census test kept that assumption [3], which Winkler notes is the naive Bayes assumption of machine learning [1].

Three refinements recur.

- **Partial agreement.** Jaro's string comparator accounts for insertions, deletions and transpositions [1], and the Jaro-Winkler version in BigMatch gives more weight to agreement in the first four characters [3]. In Winkler's Table 6, SHACKLEFORD against SHACKELFORD scores 0.970 on Jaro and 0.982 on Winkler's version, against 0.818 on edit distance [1].
- **Value frequency.** Agreement on a rare value is stronger evidence. Winkler's example is the surname Martinez, relatively rare in Minneapolis and frequent in Los Angeles [1]. The Splink documentation illustrates the idea with an earlier after-the-fact method: if pairs sharing a surname have an average match probability of 0.2 and pairs sharing "Smith" 0.1, the weight for Smiths should be adjusted down. Splink instead adds a separate term-frequency adjustment to each comparison [30].
- **One-to-one assignment.** Where a record can match at most one record in the other file, Jaro used a linear sum assignment because greedy assignment made more errors [1].

Weights can be learned without labels. Winkler showed in 1988 how to estimate them with the EM algorithm, and cautions that true matches should exceed 5% of the pairs it is applied to [1]. Splink says its models can be trained with an unsupervised approach [27]. Scale limits this. In the 2024 test, Splink's EM algorithm "could not converge in a timely manner" on the unblocked records and learned its parameters on the blocked set [3]. The Census Bureau describes its own engine as optimized for fast linking and lacking in nuance, in need of richer comparisons such as dictionary-assisted fuzzy string matching [2].

4. The clerical review band holds the uncertain pairs

The decision rule has three outcomes. A pair above the upper threshold is a match, a pair below the lower threshold is a non-match, and a pair in between is "a possible match" held for clerical review [1]. The thresholds "are determined by a priori error bounds on false matches and false nonmatches" [1].

Agencies set the band to their purpose. In the PVS, the analyst chooses both cutoffs, and because its key links one person across files, "the cutoff thresholds are set conservatively" [4]. The 2024 test set each tool's single cutoff to maximize F1 against the truth set [3]. Splink calls threshold selection "a key decision point within a linkage pipeline", a trade between false positives and false negatives [31, 33].

Clerical review also builds the truth set. In the 2020 Post-Enumeration Survey, the computer matching ran first, then clerical staff reviewed its links and searched for matches it had missed [3]. Those links became the truth set for the 2024 test. Reviewers see context the software lacks, such as household composition or whether a person moved, and can still make errors [3]. Reviewed pairs also train models: Winkler's semi-supervised method combines "a small amount (less than 0.5%) of training data from pairs in the clerical review region" with unlabeled data [1].

5. Matched pairs become entities through clustering

Pairwise decisions can contradict each other. Binette and Steorts note that most pairwise methods treat pairs as independent, "without accounting for the consequences of transitivity" [5]. Much of the literature therefore adds clustering as a second step, so that each record is assigned to its true, unknown entity [5].

One of the first methods takes the connected components of the graph of links, proposed by Monge and Elkan in 1997 [5]. Hassanzadeh and coauthors call this partitioning, or transitive closure, and show that it can produce big clusters of dissimilar records [9]. On 29 synthetic datasets it gave poor duplicate groups even against equally fast algorithms, and correlation clustering, a more sophisticated method, gave lower accuracy than some single-pass algorithms [9]. No algorithm produced perfect clusters, and the authors recommend keeping each algorithm's quantitative information alongside the clustering [9].

Thresholds shape the clusters. Splink gives each cluster a new identifier, which it calls "the ultimate output of a Splink pipeline" [32]. In its worked example, a threshold of 0.95 yields one cluster, 0.99 two and 0.999 three [32]. Its guidance says links feeding clusters require the clusters to be evaluated too [33], and that a very large cluster, say 100 records in a deduplicated dataset, can be suspicious [35].

Across many files at once, the Census Bureau describes linkage that ensures transitivity and prevents logical contradictions, after Sadinle and Fienberg (2013). It calls rigorous multi-file linkage NP-hard and beyond traditional Fellegi-Sunter methods [2]. Canonicalization, the last stage, merges each cluster into one representative record [5].

6. Registry identifiers and where they stop

A shared identifier turns a fuzzy match into an exact one where both sources carry it. The LEI is "a unique 20-character alphanumeric code", and every LEI "can represent only one entity" [18]. Each record carries a primary legal name and optional lists of other names and transliterated names [19].

The registry records its own resolution. The LEI-CDF 3.1 format defines statuses including these [19]:

- DUPLICATE: assigned to the non-surviving registration of the same legal entity, "the LEI that should no longer be used".
- LAPSED: not renewed by the next renewal date and not known by public sources to have ceased operation.
- RETIRED: the entity ceased operation without merging into another.
- ANNULLED: marked erroneous or invalid after issue.

Records also name a successor LEI where one exists, and a next renewal date by which the data must be re-certified [19].

Counts from the golden copy published at 00:00 UTC on 7 November 2026 [20]:

Registration status	LEI records	Share of all
Issued	2,001,906	57.4%
Lapsed	1,217,492	34.9%
Retired	258,744	7.4%
Duplicate	6,518	0.2%
Annulled	1,799	0.05%
All records	3,487,659	100%

Shares are computed from the counts. The remaining statuses hold a few hundred records each [20].

A lapsed LEI still identifies its entity. GLEIF wrote in March 2017, when some 30% of LEIs had lapsed, that a lapsed LEI "remains a valid LEI" whose entity is behind with its renewal [21]. A match to an LEI therefore carries its status and renewal date with it [19].

The LEI stops at people. The LEI Regulatory Oversight Committee states that natural persons acting in a private capacity, and professionals acting as employees, are not eligible, and quotes ISO 17442 as excluding natural persons [22].

Crosswalks are matching work too. GLEIF certifies organizations that map their own identifiers to the LEI and publishes relationship files for BIC, ISIN, MIC, S&P Capital IQ and OpenCorporates identifiers [23]. The Census Bureau's own person key is itself the output of probabilistic linkage against reference files, one of them holding more than 1.2 billion observations, and then links that person across every file processed [4].

7. Language-model matchers in published tests

Peeters, Steiner and Bizer tested three hosted models from one provider and three open-weight models as entity matchers (EDBT 2025) [10]. They used six public benchmarks, four of product offers and two bibliographic, each test set cut to about 1,250 pairs, with 193 to 259 positives per set [10]. On those benchmarks:

- **Zero-shot.** The top hosted model reached 89% F1 or higher on five of the six datasets without task-specific training data [10].
- **Against fine-tuned matchers.** It scored above the best fine-tuned baseline on three datasets, by 2.65 to 4.71 F1 points, and below it on three, by 3.69, 4.49 and 0.73 points [10].
- **Unseen entities.** Fine-tuned matchers applied to unseen products lost 36 to 56 F1 points (Ditto) and 22 to 61 points (RoBERTa), and every model tested scored at least 8 points above the best transferred matcher [10].
- **Prompts.** "There is no single best prompt". The best prompt depends on the model and dataset [10].
- **Open-weight models.** These reached F1 similar to hosted models given a small amount of training data or matching rules [10].
- **Cost.** At prices as of August 2024, in-context learning and rule-based prompting cost 1.5 to 470 times as much as zero-shot prompting of a low-cost hosted model [10].

The authors recommend fine-tuned smaller models where few entities are unseen and training data exists, and language models where many entities are unseen and labels are costly [10]. Three further results set limits.

Benchmark difficulty. Papadakis, Kirielle, Christen and Palpanas (2023, published at ICDE 2024) assessed 13 established datasets and found that most "pose rather easy classification tasks" [12]. Their bar for a challenging dataset is a non-linear boost and a learning-based margin both above 5%, ideally 10%. Four datasets met it [12]. The structured DBLP-ACM and DBLP-Scholar sets used in [10] were among those many algorithms solve almost perfectly [12]. A 2026 study added synthetic datasets of fictional items because public benchmarks may have been seen in pretraining [14].

Cost scaling. Fan and coauthors work through a table of 1,000 records needing about 500,000 pair predictions. At a hosted model's December 2023 price of $0.01 per 1,000 tokens, standard prompting with three demonstrations would cost about $1,800 [11]. Batching questions into one prompt gave a 4x to 7x cost saving with higher and more stable accuracy [11]. A 2025 method that asks the model to cluster sets of records reports up to five times fewer API calls on nine datasets [15].

Explanations. Teofili and coauthors (May 2026) tested self-explanations across three models and ten datasets and found them "often unstable, weakly faithful, and poorly aligned with counterfactual evidence" [14]. Post-hoc explanation methods were more trustworthy and carried a prohibitive computational cost [14].

For company records, the one case among these sources uses a fine-tuned model. Ditto matched employer tables of 789,409 internal and 412,418 public records (788,094 and 62,511 after deduplication) on name, address, city, state, zip code and phone [13]. From 20,000 labeled candidate pairs, 39% positive and sampled toward high-similarity pairs, it reached 96.53 F1 on a held-out split [13]. The figure applies to those sampled pairs only.

8. Measuring the result

Every measure needs a truth set, and human-labeled results "rarely number beyond several thousand samples" [8]. Entity resolution metrics leave out true negatives, because non-matching pairs grow with the square of the records and metrics built on them, such as the false positive rate, give overly optimistic results for large databases [8]. Splink's documentation makes the same point about accuracy, which looks high by "simply assuming the majority class" [34].

The main measures [7, 8]:

Measure	What it rewards	Note in the sources
Pairwise precision, recall and F1	Correct record pairs	The most used. Mismatching two singletons barely moves it, against joining or splitting large clusters
Cluster F1	Exactly reproduced clusters	One wrong record makes a whole cluster count as wrong
Purity	Clusters free of foreign records	One record per cluster scores perfect purity
Variation of Information	Information lost and gained between result and truth	A special case of generalized merge distance
Generalized merge distance	The cost of splits and merges needed to reach the truth	Configurable. Pairwise F1 can be computed from it

The measures conflict. Menestrina, Whang and Garcia-Molina count a conflict when two measures prefer opposite results from the same pair of algorithms [7]. On synthetic data, one measure pair conflicted on 52 of 100 result pairs, the average conflict probability across all result pairs was 9.5%, and errors of a single type produced no conflicts [7]. On real hotel data, two measures ranked the same two algorithms in opposite order [7]. In their worked example, one result finds 15 correct pairs and another 8, yet the second resolves two whole entities correctly [7]. They conclude that "evaluations of ER algorithms need to consider multiple measures", which is seldom done [7]. Barnes recommends pairwise F1 at a minimum, plus a cluster metric and generalized merge distance [8].

The Census Bureau's 2024 test shows how far tools differ on the same records. Three tools deduplicated the same records on the same six fields, fastLink under its looser blocking scheme, each at its F1-optimal cutoff [3]:

Test set	Tool	Seconds	Precision	Recall	F1
United States and D.C., 1,167,705 records, 3,941 true links	BigMatch	38	0.8733	0.7161	0.7869
	fastLink	44,161	0.9636	0.4568	0.6198
	Splink	40	0.7325	0.6336	0.6795
Puerto Rico, 33,254 records, 1,125 true links	BigMatch	1	0.9522	0.8277	0.8849
	fastLink	7,523	0.8905	0.6396	0.7445
	Splink	11	0.71	0.8587	0.7773

The tool with the top national precision had the lowest recall, and Splink had the top recall in Puerto Rico [3]. The authors flag possible "bias in favor of BigMatch", because the first round of deduplication in the 2020 Post-Enumeration Survey used software that incorporates it [3]. The Census research page cites a 2025 paper by the same four authors (Arthun, Gilary, McGinnis and Zamora) when it says the Bureau's software performs better than open-source competitors in general [2]. Splink's claim of linking a million records on a laptop in about a minute is the project's own [27], and the 2024 test timed a different task [3]. GUILD adds that accuracy should be reported for important subgroups, such as infants and the elderly, for whom it may vary [16].

9. Linkage error carries into the analysis

Doidge and Harron define linkage error as missed links between records of one person or false links between unrelated records [17]. It can act as missing data, measurement error, misclassification or unrepresentative sampling. It can also split or merge people, as when two hospital admissions count as one person admitted twice if linked and two people admitted once if not [17]. In Winkler's illustration, regression relationships nearly disappear as matching error rises [1]. GUILD warns that linkage error "disproportionately affects disadvantaged groups" [16].

The remedy the sources give is disclosure at the record level. GUILD asks data linkers to provide record-level indicators such as the pass ID or match scores, to report how records were blocked, and to share the optimum threshold and, where possible, all possible links above it for each record [16]. The PVS lists links by pass with their scores and the records matched from each file, so an analyst can check the lowest-scoring

links in the loosest pass [4]. Winkler recommends that corrected register entries carry a dated certification status code [1]. The LEI index keeps duplicate registrations under their status [19, 20], and the SEC keeps former names with dates [26].

Implications

For data engineers building a pipeline. Plan blocking as several strict passes on independent fields, and measure pair completeness and reduction ratio on labeled data [4, 6, 28]. Record cleaning rules as part of the method [16]. Choose clustering deliberately: transitive closure, the common approach in early work, gave poor groups in a published test on synthetic data [9]. Keep each pair's score, pass and threshold with the output [16]. Test language-model matchers on data like the target, since public benchmarks are often easy [12], and treat their verdicts and explanations as separate outputs [14].

For buyers commissioning or licensing a dataset whose entities must resolve. Ask for the truth set behind any accuracy figure, its size and labeling, the threshold, and at least one pairwise and one cluster measure [7, 8]. Ask how blocking was done, since it caps recall [6, 28]. Ask whether source records survive resolution, and whether identifiers such as the LEI are read with their status and dates [19, 21].

For analysts using linked data. Linkage error can split or merge the people being counted [17]. Record-level scores and thresholds let analysts adjust for it [16].

Rillor scopes, builds and maintains custom datasets to a client's requirement and reconciles identities without erasing the source record. Each one comes with its schema, records with resolved identities, source provenance and lineage.

Limits of this analysis

- Winkler's overview dates from 2006. Newer sources [2, 3, 5, 6, 10, 14] cover current practice, and some figures, such as the census blocking result, are older.
- The clustering comparison [9] and most measure-conflict results [7] use synthetic data.
- The language-model results [10] cover six public benchmarks with test sets of about 1,250 pairs, several of them easy by the measures in [12]. Prices are from August 2024 [10] and December 2023 [11].
- The Census tool comparison [3] may favor the Bureau's own software, as its authors state, and its clerical truth set can contain errors.
- Splink's speed, scale and user claims are self-reported [27].
- No source gives cross-industry error rates for company matching, LEI coverage of all companies, a standard size for the review band or a head-to-head of language models and Fellegi-Sunter tools at registry scale.
- The 1969 Fellegi and Sunter paper was not opened. The model is described through [1, 3, 29].
- This report describes what the sources show. It gives no legal advice on the use of data.

Sources

1. Overview of Record Linkage and Current Research Directions, Statistics #2006-2 (William E. Winkler). U.S. Census Bureau, 8 February 2006. https://www.census.gov/content/dam/Census/library/working-papers/2006/adrm/rrs2006-02.pdf

2. Record Linkage and Machine Learning, research topic page. U.S. Census Bureau, last revised 29 October 2026, viewed 7 November 2026. https://www.census.gov/topics/research/stat-research/expertise/record-linkage.html

3. Comparing Open-Source Record Linkage Software to Existing Approaches: A Case Study Using Census Data (Arthun, Gilary, McGinnis and Zamora, U.S. Census Bureau). Zenodo, 21 October 2024. https://zenodo.org/records/13961108

4. The Person Identification Validation System (PVS), CARRA Working Paper #2014-01 (Wagner and Layne). U.S. Census Bureau, 1 July 2014. https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf

5. (Almost) All of Entity Resolution (Binette and Steorts). Science Advances 8(12), 25 March 2022, arXiv version. https://arxiv.org/abs/2008.04443

6. A Survey of Blocking and Filtering Techniques for Entity Resolution (Papadakis, Skoutas, Thanos and Palpanas). arXiv, version of 21 August 2020. Published in ACM Computing Surveys 53(2), 2020. https://arxiv.org/abs/1905.06167

7. Evaluating Entity Resolution Results (Menestrina, Whang and Garcia-Molina). Proceedings of the VLDB Endowment 3(1), September 2010. https://vldb.org/pvldb/vol3/R18.pdf

8. A Practioner's Guide to Evaluating Entity Resolution Results, title as published (Barnes). arXiv, 14 September 2015. https://arxiv.org/abs/1509.04238

9. Framework for Evaluating Clustering Algorithms in Duplicate Detection (Hassanzadeh, Chiang, Lee and Miller). Proceedings of the VLDB Endowment 2(1), August 2009. https://www.vldb.org/pvldb/vol2/vldb09-1025.pdf

10. Entity Matching using Large Language Models (Peeters, Steiner and Bizer). EDBT 2025, arXiv version of 18 October 2024. https://arxiv.org/abs/2310.11244

11. Cost-Effective In-Context Learning for Entity Resolution: A Design Space Exploration (Fan and coauthors). arXiv, 7 December 2023. https://arxiv.org/abs/2312.03987

12. A Critical Re-evaluation of Benchmark Datasets for (Deep) Learning-Based Matching Algorithms (Papadakis, Kirielle, Christen and Palpanas). arXiv, version of 13 November 2023. A revised version appeared at IEEE ICDE 2024 as "A Critical Re-evaluation of Record Linkage Benchmarks for Learning-Based Matching Algorithms". https://arxiv.org/abs/2307.01231

13. Deep Entity Matching with Pre-Trained Language Models (Li, Li, Suhara, Doan and Tan). Proceedings of the VLDB Endowment 14(1), 2020, arXiv version. https://arxiv.org/abs/2004.00584

14. Can we trust LLM Self-Explanations for Entity Resolution? (Teofili, Firmani, Koudas, Merialdo and Srivastava). arXiv, 31 May 2026. https://arxiv.org/abs/2606.01210

15. In-context Clustering-based Entity Resolution with Large Language Models: A Design Space Exploration (Fu and coauthors). arXiv, 3 June 2025. https://arxiv.org/abs/2506.02509

16. GUILD: GUidance for Information about Linking Data sets (Gilbert and coauthors). Journal of Public Health 40(1), published online 28 March 2017. https://pmc.ncbi.nlm.nih.gov/articles/PMC5896589/

17. Reflections on modern methods: linkage error bias (Doidge and Harron). International Journal of Epidemiology 48(6), published online 21 October 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC7020770/

18. ISO 17442: The Global Standard. Global Legal Entity Identifier Foundation, viewed 24 October 2026. https://www.gleif.org/en/about-lei/iso-17442-the-lei-code-structure

19. Level 1 Data: LEI-CDF Format 3.1. Global Legal Entity Identifier Foundation, viewed 24 October 2026. https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-1-data-lei-cdf-3-1-format

20. GLEIF API, LEI records filtered by registration.status, golden copy of 7 November 2026 00:00 UTC. Global Legal Entity Identifier Foundation, viewed 7 November 2026. https://api.gleif.org/api/v1/lei-records?filter%5Bregistration.status%5D=ISSUED&page%5Bsize%5D=1

21. The Power of Transparency: A Closer Look at LEI Renewal Rates. Global Legal Entity Identifier Foundation, 2 March 2017. https://www.gleif.org/en/newsroom/blog/the-power-of-transparency-a-closer-look-at-lei-renewal-rates

22. Statement on Individuals Acting in a Business Capacity. LEI Regulatory Oversight Committee, 30 September 2015. https://www.leiroc.org/publications/gls/lou_20150930-1.pdf

23. LEI Mapping. Global Legal Entity Identifier Foundation, viewed 24 October 2026. https://www.gleif.org/en/lei-data/lei-mapping

24. ISO 20275: Entity Legal Forms Code List. Global Legal Entity Identifier Foundation, viewed 24 October 2026. https://www.gleif.org/en/about-lei/code-lists/iso-20275-entity-legal-forms-code-list

25. EDGAR Application Programming Interfaces. U.S. Securities and Exchange Commission, viewed 24 October 2026. https://www.sec.gov/search-filings/edgar-application-programming-interfaces

26. Submissions record for CIK 0001326801. U.S. Securities and Exchange Commission, viewed 24 October 2026. https://data.sec.gov/submissions/CIK0001326801.json

27. Splink documentation, home page. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/index.html

28. Splink: What are Blocking Rules? UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/blocking/blocking_rules.html

29. Splink: The Fellegi-Sunter model. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/theory/fellegi_sunter.html

30. Splink: Term-Frequency Adjustments. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/comparisons/term-frequency.html

31. Splink: Types of Record Linkage. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/theory/probabilistic_vs_deterministic.html

32. Splink: Linked data as graphs. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/theory/linked_data_as_graphs.html

33. Splink: Edge Evaluation. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/evaluation/edge_overview.html

34. Splink: Edge Metrics. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/evaluation/edge_metrics.html

35. Splink: Cluster Evaluation. UK Ministry of Justice Analytical Services, viewed 24 October 2026. https://moj-analytical-services.github.io/splink/topic_guides/evaluation/clusters/overview.html

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How entity resolution matches records across sources. 7 November 2026. https://rillor.com/insights/how-entity-resolution-matches-records

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data