Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Entity Resolution Blocking Recall: Audit for Hidden Matches

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-27 · entity resolution blocking recall · entity resolution blocking · record linkage blocking · blocking recall · pairs completeness · reduction ratio · entity resolution candidate generation · missed matches · blocking audit

Original article: https://reliablecontext.com/articles/entity-resolution-blocking-audit-missed-matches



In this article

1. Executive Summary
2. Introduction and Background
3. Defining the Audit Frame
4. Record Linkage Blocking Methods and How They Hide Pairs
5. Can the Resolver Expose Its Candidate Pairs?
6. Adjudicated Reference Pairs and the Missing-Label Problem
7. A Synthetic Cross-Source Fixture (Hypothetical Example)
8. Stratifying Omissions and Rerunning After a Source Change
9. Decision Rule and the Audit Worksheet
10. Data Analysis and Evidence
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

Entity resolution blocking is the candidate-generation stage that decides which record pairs a matcher will ever compare. A true match that never becomes a candidate cannot be recovered by any later scoring or clustering step (recordlinkage.readthedocs.io). The survey literature separates this stage from matching: Papadakis et al. (2020) restrict their survey to candidate selection and describe blocking as typically independent of the matching step (arxiv.org). This report turns that separation into an acceptance test with two numbers, both computed before any matcher score is read. **Candidate recall** (the literature calls it pairs completeness) is the adjudicated true pairs present in the candidate set divided by all adjudicated true pairs (users.cecs.anu.edu.au). **Reduction ratio** is one minus the candidate comparisons divided by the eligible comparisons, with the eligible set declared as cross-source or within-source (users.cecs.anu.edu.au).

A synthetic fixture (Hypothetical Example) shows why both numbers are needed. Two blocking rules over two ten-record sources, with 100 eligible cross-source comparisons and nine adjudicated true pairs, produced 11 candidates, captured 8 of 9 true pairs (candidate recall **0.889**) and reached a reduction ratio of **0.89**. After one upstream date-format change, the same rules produced 9 candidates, captured 6 of 9 (**0.667**), and the reduction ratio rose to **0.91**. An efficiency number that improves while recall falls is the signature of silent omission, and only a labeled-pair comparison exposes it.

The published evidence is method-specific and should not be generalized. Papadakis et al. (2022) tuned 14 filters and 4 baselines on 10 real-world datasets toward a pair-completeness target (arxiv.org), and a separate benchmark applied default parameters instead. A Washington State Department of Health study of 51,401 verified marriage and death certificate links found exact date-of-birth blocking retained 96.7%, 95.4% and 92.6% of true links across White/Non-Hispanic, Non-White and/or Hispanic, and Missing Race/Ethnicity groups (doh.wa.gov). A Hong Kong study found that blocking on non-standardised romanisation reached 68.8% recall while standardised schemes exceeded 95% (ijpds.org).

Three decisions follow. **Accept** when the candidate pairs are observable, the reference pairs were adjudicated independently of the candidate set, and recall and every stratum meet floors fixed before the test. **Revise candidate coverage** when named true pairs or strata are missing. **Hold, with the unknown declared**, when the system does not expose its candidate pairs: that test is unavailable, not passed. Labels drawn only from a system's own output understate omissions (www.cs.cornell.edu).

Introduction and Background

Entity resolution (ER) decides which records describe the same real-world thing. Record linkage is the older name for the same problem across files, and the survey literature treats blocking as the way to avoid comparing every pair, since the naive comparison space grows quadratically (users.cecs.anu.edu.au) (researchportalplus.anu.edu.au). A 2024 explainer from a human-rights organisation, which describes the pipeline for a general audience, likewise places "preprocessing and blocking" first and links to the Papadakis survey for it (www.amnesty.org). Practitioner explainers state the tradeoff in general terms: blocks that are too strict hide true matches, and the tradeoff is precision versus recall at the candidate stage (tilores.io). Others note that poor blocking produces false negatives at scale (www.moderndata101.com). What these treatments leave open is a procedure that a reviewer can run on someone else's resolver, with labeled pairs, before arguing about scores.

This report is that procedure. It is written for context engineers and identity stewards who must approve a candidate stage for a defined record population, and who need to decide whether candidate coverage is adequate before interpreting matcher scores or final clusters. The test is observable at the input and output of any resolver: it needs the records, the candidate pairs the resolver emits, and a set of independently adjudicated true pairs. It does not need the resolver's internals, and nothing here describes any particular product's mechanism.

The publisher, Reliable Context, is not a record-linkage library, a master data management (MDM) suite or a data-quality suite; readers who need a blocking library should look at the open-source tools in Table 1. The audit below is independent of any resolver, including the publisher's. Downstream cluster evaluation (pairwise F1, bridge records) is a separate question, and this report deliberately stops before it.

Public search interest in the exact phrasing is small. An internal keyword-tool snapshot, not reproducible from any public source, retrieved Google Ads volume estimates through the DataForSEO API on 2026-09-27; it showed 720 monthly searches for "entity resolution", 110 for "record linkage" and 10 for "entity resolution blocking" in August 2026, with no reported estimate for "blocking recall" phrasing. The API's field is documented as a monthly average search volume rate based on Google Ads data (docs.dataforseo.com), so these are advertising estimates, not traffic forecasts, and no trend is claimed.

Defining the Audit Frame

Terms

The audit needs five defined terms and one adjudication rule, stated before any data are touched.

- **Record population:** the set of records in scope, with source names and extraction dates. The audit result applies to this population only.
- **Eligible comparison:** a pair the resolver is allowed to consider. For linking two sources A and B it is the Cartesian product, with m times n pairs; for deduplicating one source of n records it is n choose 2. The reduction-ratio documentation of one blocking package gives the two formulas separately for exactly this reason (blockingpy.readthedocs.io), and the Python Record Linkage Toolkit's index step returns candidate pairs as a MultiIndex of record-label pairs (recordlinkage.readthedocs.io).
- **True-match pair:** two records an adjudicator judged to describe the same real-world thing, using evidence outside the resolver's own output.
- **Candidate pair:** an eligible pair the candidate stage passed to the matcher, whether or not it later scores as a match.
- **Adjudication frame:** the written inclusion rules, the adjudicator, the evidence basis and the dates, so that a missing pair can be classified as a true omission rather than a labeling disagreement.

The two measures

Christen's survey of indexing techniques defines pairs completeness as the true matched candidate pairs generated divided by the total number of true matched pairs (users.cecs.anu.edu.au), and reduction ratio as the fraction of the comparison space removed (users.cecs.anu.edu.au). Papadakis et al. give the same pair-completeness definition and call it an optimistic estimate of recall, because it presumes a perfect matcher: a pair is counted as detected as long as the two records share a block (arxiv.org). This report uses **candidate recall** as the plain-language name for that quantity:

candidate recall = adjudicated true pairs present in the candidate set / all adjudicated true pairs

reduction ratio = 1 minus (candidate comparisons / all eligible comparisons)

Reduction ratio says nothing about whether surviving candidates are true matches (users.cecs.anu.edu.au). The trade between the two is structural: the more comparisons the block collection contains, the more duplicates it can detect, at lower efficiency (users.cecs.anu.edu.au), and Christen compares that trade to the precision-recall trade (users.cecs.anu.edu.au).

Why a recall ceiling matters

Pairs removed in the indexing step are treated as non-matches without comparison (spiral.imperial.ac.uk), and Papadakis et al. (2022) state the consequence directly: the recall of the filtering step determines the overall recall (arxiv.org). Older work makes the same point in plainer terms: an erroneous blocking-key value can place a record in the wrong block, leading to missing true matches (users.cecs.anu.edu.au), and nothing in the later resolution step brings the pair back. The consequence for approval is that a high final-cluster score can coexist with unseen losses: the matcher can only be as complete as its candidates.

Record Linkage Blocking Methods and How They Hide Pairs

Blocking is a family of methods, and each hides pairs through a different mechanism. This section answers "what is blocking in entity resolution" by mechanism, because the mechanism predicts which strata to inspect.

Hash and key-based blocking

Traditional blocking assigns each record a blocking key value and compares only records in the same block; Christen traces the technique to the 1960s (users.cecs.anu.edu.au), and a 2007 survey defines it as subdividing files into mutually exclusive subsets on the assumption that no matches occur across blocks (archive.nyu.edu). Its failure mode is exact disagreement: any error in a key field can send a true match to another block (users.cecs.anu.edu.au). Steorts and colleagues note that, because fields can be unreliable in many applications, blocking may miss large proportions of matches (arxiv.org). A second survey adds that errors in the blocking step can place entries in the wrong buckets and cause missed matches (archive.nyu.edu), and Christen notes that choosing key fields so that as many true pairs as possible become candidates is a question separate from the choice of technique (users.cecs.anu.edu.au).

Window, similarity and nearest-neighbour methods

Sorted neighbourhood sorts by a key and compares records inside a sliding window (users.cecs.anu.edu.au) (hpi.de); a 2007 survey notes that no single sort key can place all matching records together, which motivates multiple passes (archive.nyu.edu). Q-gram indexing places records with similar, not only identical, key values in the same block (users.cecs.anu.edu.au), suffix-array indexing indexes suffixes of key values (users.cecs.anu.edu.au), and canopy clustering uses a cheap similarity to form overlapping subsets (www.cs.cmu.edu) (users.cecs.anu.edu.au). Papadakis et al. (2020) classify block building along several dimensions and observe that blocking is an inherently approximate procedure (arxiv.org). Block-processing steps between building and matching then prune comparisons, and the survey positions them between blocking and matching.

Rule unions

Practical tools combine rules. Splink produces all comparisons that satisfy at least one blocking rule (moj-analytical-services.github.io), the Python Record Linkage Toolkit takes the union when several indexing algorithms are added (recordlinkage.readthedocs.io), and Splink's guidance concedes that a single rule would miss a true match with a typo, which is why several rules are listed (moj-analytical-services.github.io). The same guidance states the dual aim of blocking rules: capture as many true matches as possible while limiting the comparisons generated (moj-analytical-services.github.io). The fixture below uses a two-rule union for the same reason, and the audit evaluates the union, not each rule alone.

Can the Resolver Expose Its Candidate Pairs?

The audit's first gate is observability. If the evaluated system does not emit or allow export of the pairs it passed to matching, candidate recall cannot be computed, and the result is **unavailable**, not passed. Stating a recall figure from the matcher's final links would answer a different question, because the final links are a subset of the candidates.

Table 1 summarizes what the official documentation of six well-known open-source tools says about obtaining candidate pairs. It records documentation statements only, as of October 2026; it is not a ranking, and a tool absent from the table has not been judged.

Tool (official docs)	Documented way to obtain candidate pairs	Documented candidate-stage measures	Gap for a recall audit
Python Record Linkage Toolkit	The index step returns a MultiIndex of record pairs (recordlinkage.readthedocs.io)	Reduction ratio accepts the pair count or the MultiIndex itself (recordlinkage.readthedocs.io); recall is TP/(TP+FN) (recordlinkage.readthedocs.io)	No named candidate-recall function; passing the candidate index as predicted links is an inference of this report
Splink	A method materialises the blocked pairs that prediction would score (moj-analytical-services.github.io)	Per-rule comparison counts, estimated from a 5% sample by default (moj-analytical-services.github.io); a chart reports reduction against the Cartesian product (moj-analytical-services.github.io)	Counts are of pairs generated, not of true matches captured
Dedupe	A pairs method yields records sharing a fingerprint (docs.dedupe.io)	A training parameter sets the share of true dupe pairs the blocking must cover	Lowering it leaves true dupes that will never be compared (docs.dedupe.io)
py_entitymatching (Magellan)	Blockers return the surviving candidate set as a DataFrame (anhaidgroup.github.io)	A debugger reports likely true matches rejected by the blocker (anhaidgroup.github.io)	Output is drawn from the Cartesian product minus the candidate set (anhaidgroup.github.io); it ranks suspects, it does not measure recall
pyJedAI	Every pipeline step has an export method returning pairs as a dataframe (pyjedai.readthedocs.io)	Tutorial reports recall against a ground truth after each step	Purging steps discard blocks over a comparison limit (pyjedai.readthedocs.io), so each step needs its own check
Zingg	Documentation read for this report describes per-block counts and samples	A verify-blocking phase shows counts per block (docs.zingg.ai)	No full pair export or recall measure found in the verify-blocking documentation page read; other Zingg pages were not checked

Two cautions apply to the table. A tool that exposes pairs in a library is not the same as a deployed resolver exposing them, so the question must be put to each deployment. And exposure is necessary, not sufficient: the Splink counts are estimates unless the sampling proportion is raised (moj-analytical-services.github.io), so an auditor must check whether a reported candidate count is exact or sampled before using it as the numerator of a reduction ratio.

Adjudicated Reference Pairs and the Missing-Label Problem

Candidate recall is only as credible as its denominator. The denominator is the set of all true pairs, but nobody holds that set; they hold adjudicated pairs, and the way those were chosen decides whether the measured recall is meaningful.

Labels that came from the system under test

Information retrieval met this problem decades ago. Zobel's analysis of pooled relevance judgments found that measured relative performance appeared reliable but recall was overestimated, because relevant documents that no submitted system retrieved were never judged (www.cs.cornell.edu), since unjudged documents are assumed not relevant (www.cs.cornell.edu). Techniques that did not contribute to the pool were at risk of having their effectiveness underestimated (www.cs.cornell.edu), and later work showed that common evaluation measures are not robust to substantially incomplete judgments (www.nist.gov). For ER the analogue is exact: if the reference pairs were produced by running the resolver (or its predecessor) and reviewing its output, the pairs the candidate stage hides are the pairs least likely to be labeled. Marchant and Rubinstein warn that evaluation which filters candidates with blocking injects hidden bias into estimates (arxiv.org), and a dense-blocking benchmark paper observes that datasets built from existing matching benchmarks may carry evaluation bias for blocking (arxiv.org).

A second circularity is label provenance. Binette and Steorts note that benchmark results depend on the reliability of the identifiers used for evaluation (arxiv.org), and that some public identifiers were obtained as the result of extensive record linkage efforts involving expert clerical review of the data (arxiv.org). A reference set labeled by a prior linkage run inherits that run's blind spots. Reviewers of linkage studies make the same point about missed matches when the outcome is unknown: it is difficult to know which matches have been missed (www.ncbi.nlm.nih.gov). The inclusion rule should therefore name the evidence basis for each pair (a third document, a registry cross-reference, a manual review of unlinked records) and flag any pair whose evidence is another linkage output.

Evaluation tools constrain what can be claimed

One blocking package's own paper illustrates the boundary: its evaluation is performed only for records available in the supplied ground truth, so the reported recall holds for the labeled subset (arxiv.org). The general rule is that a recall figure carries the scope of its labels. A related point: accuracy-style measures are not meaningful in linkage because of class imbalance, and precision and recall are used instead, conditional on ground truth being available (spiral.imperial.ac.uk).

When a confidence interval is legitimate

A confidence interval describes sampling variability, so it requires a probability sample: true pairs drawn at random from a frame that was built independently of the candidate set, with known inclusion probabilities. A convenience set, a constructed fixture, or a list of known problem cases supports counts and named omissions but not an interval. When a probability sample exists, the Wilson score interval is the sensible default: Brown, Cai and DasGupta recommend the Wilson interval for small n (asaip.psu.edu), and Agresti and Coull describe it as inverting the score test with the null standard error, giving coverage near nominal even for very small samples (math.unm.edu). Where extreme class imbalance makes uniform sampling of candidate matches expensive, importance-sampling designs exist; one reports up to 83% fewer labels without loss of estimate accuracy under its stated conditions (arxiv.org).

(Hypothetical Example) Suppose a frame of 2,400 independently adjudicated true pairs and a simple random sample of 120, of which 108 appear in the candidate set. The sample recall is 108/120 = 0.900, and the 95% Wilson interval is 0.833 to 0.942, computed with the standard formula. Had all 40 of a 40-pair sample been present, the interval would still run from 0.912 to 1.000, so zero observed misses is not a claim of perfection. The fixture below is not a probability sample, so it receives no interval.

A Synthetic Cross-Source Fixture (Hypothetical Example)

The fixture is small enough to verify by hand and built to contain the omission patterns that real populations produce: transliterated or script-changed names, a missing identifier, shared addresses, a changed name, a swapped name order, and one deliberately hard pair. All values are invented. Source A is dated 2026-09-01 and Source B 2026-09-15.

Source A:

```
rec,given,family,dob,address
A1,Mohammed,Al-Rashid,1984-03-12,14 Quay Rd
A2,Elena,Petrova,1979-11-02,7 Birch Ln
A3,Maria,Keller,1990-06-21,22 Mill St
A4,Sergei,Ivanov,1971-01-30,3 Oak Ct
A5,John,Smith,1965-09-14,9 Elm Ave
A6,Wei,Zhang,1988-04-05,41 Pine Rd
A7,Priya,Nair,1995-12-19,5 Cedar Way
A8,Thomas,Becker,1958-07-08,12 Lake Dr
A9,Anna,Novak,2001-02-27,30 Hill Rd
A10,David,Smith,1992-10-10,9 Elm Ave
```

Source B:

```
rec,given,family,dob,address
B1,Muhammad,Alrashid,1984-03-12,"Flat 2, 14 Quay Road"
B2,Yelena,Petrova,,7 Birch Lane
B3,Maria,Hoffmann,1990-06-21,22 Mill Street
B4,Сергей,Иванов,1971-01-30,3 Oak Court
B5,John,Smith,1965-09-14,9 Elm Avenue
B6,Zhang,Wei,1988-04-05,41 Pine Road
B7,Pria,Nayar,1995-12-19,77 Ridge St
B8,Tomas,Becker,1985-07-08,12 Lake Drive
B9,Ana,Novak,2001-02-27,PO Box 8
B10,Robert,Jones,1980-05-05,9 Elm Avenue
```

Note that in B6 the family field holds "Wei" and the given field holds "Zhang" (swapped order). The adjudicated reference table is below. Adjudication used evidence outside both exports (a paper form carrying a document number held in neither source), and the adjudicator is independent of the blocking configuration.

```
pair,basis,stratum_script,stratum_dob,stratum_age_2026
A1-B1,same document no.,Latin,present,40-59
```

```
A2-B2,same document no.,Latin,missing in B,40-59
A3-B3,same document no. (name change),Latin,present,under 40
A4-B4,same document no. (Cyrillic in B),Cyrillic,present,40-59
A5-B5,same document no.,Latin,present,60+
A6-B6,same document no. (name order swapped),Latin,present,under 40
A7-B7,paper form cross-reference (hard pair),Latin,present,under 40
A8-B8,same document no. (birth year typo in B),Latin,present,60+
A9-B9,same document no.,Latin,present,under 40
```

The records A10 and B10 have no true partner. The blocking configuration is the union of two rules: **Rule 1** matches on the first four ASCII-folded letters of the family name plus birth year, and requires a non-empty year; **Rule 2** matches on the normalised address, expanding Road, Lane, Street, Avenue, Court and Drive to standard abbreviations. The eligible set is **cross-source**: 10 by 10 gives 100 comparisons. Within-source pairs (45 in each source) are outside the frame, so dedupe-style counts are not mixed in.

Table 2 gives the arithmetic. Each figure was computed by script and re-derived by hand.

Run	Candidates	True pairs captured	Candidate recall	Reduction ratio	Pairs quality
Rule 1 only	3	3 of 9	3/9 = 0.333	1 − 3/100 = 0.970	3/3 = 1.000
Rule 2 only	9	6 of 9	6/9 = 0.667	1 − 9/100 = 0.910	6/9 = 0.667
Union, baseline	**11**	**8 of 9**	**8/9 = 0.889**	**1 − 11/100 = 0.890**	8/11 = 0.727
Union, after Source B date-format change	**9**	**6 of 9**	**6/9 = 0.667**	**1 − 9/100 = 0.910**	6/9 = 0.667

The baseline union misses one pair, **A7-B7**: the surname differs (Nair against Nayar) so Rule 1 fails, and the address changed so Rule 2 fails. No downstream matcher could recover it, and its adjudication basis is the only reason it appears in the denominator. Rule 1 alone looks excellent on reduction ratio (0.970) and poor on recall (0.333), which illustrates why efficiency must never be reported alone. The three false candidates (A10-B10, A10-B5, A5-B10) all come from the shared address "9 Elm Ave". They lower pairs quality and cost matcher effort but hide nothing. Had the audit chosen the combined 20-record deduplication frame instead, with 190 eligible pairs, the same 11 candidates would give a reduction ratio of 0.942, a reminder that the denominator must be declared.

Stratifying Omissions and Rerunning After a Source Change

Strata

An overall recall of 0.889 hides where the loss falls. The audit stratifies adjudicated pairs by source pair, script, identifier missingness and age, then reports numerator and denominator in every cell. The fixture has a single source pair (A by B), so that stratum is trivial here but mandatory in real populations with three or more

sources.

The literature gives reasons to expect uneven loss. A Washington State Department of Health report tested blocking on 51,401 verified marriage-to-death links and found exact date-of-birth blocking retained 96.7%, 95.4% and 92.6% across the White/Non-Hispanic, Non-White and/or Hispanic, and Missing Race/Ethnicity groups, while fuzzy date-of-birth blocking at Hamming distance 1 raised these to 99.1%, 98.6% and 96.5% (doh.wa.gov). The report's stated mechanism is that overly strict blocking gives low recall for pairs with higher rates of inexactness (doh.wa.gov). An evaluation framework for blocking defines group-level pair completeness over pairs in which at least one entity belongs to the group (arxiv.org), and finds that the best overall method is not necessarily the one with the lowest disparity (arxiv.org); its propagation test assumes a hypothetical perfect matcher so that only blocker bias remains.

Script and name conventions matter independently. The International Civil Aviation Organization's travel-document standard notes that a Latin rendering of an Arabic-script name is one of many possibilities (www.icao.int), and that diacritical marks are not permitted in the machine-readable zone (www.icao.int). A Hong Kong study of 771 student names measured blocking recall against a known Chinese surname: standardised Jyutping and Pinyin exceeded 95%, while the non-standardised government romanisation reached 68.8% (ijpds.org). Phonetic codes group surnames that sound the same but are spelled differently, like Smith and Smyth (www.archives.gov), but they do not remove the problem: an economic-history study found that, for the Ferrie (1996) method with its uncommon-name restriction, Soundex cleaning cut the match rate from 33% (uncleaned names) to 20% by creating more common name strings, while the Abramitzky et al. (2014) method showed match rates of 39 to 42 percent (www.nber.org).

Missingness and age show similar patterns. Harron and colleagues found sex and date-of-birth errors or missing values in 0.11% of hospital records but postcode errors or missing values in 53% (bmcmedresmethodol.biomedcentral.com), so a blocking rule that depends on postcode inherits a very different failure rate from one that depends on date of birth. Deterministic exact matching linked 42% of baby records to a maternal record, against 98% with probabilistic linkage (pmc.ncbi.nlm.nih.gov). A review of data-linkage studies reported that age, sex and race were unevenly distributed between matched and unmatched records in 72%, 50% and 64% of studies respectively (www.ncbi.nlm.nih.gov), with a quantified missing-identifier example of 21.4% against 3.2% by maternal age group (www.ncbi.nlm.nih.gov). Linkage quality was worse for younger individuals in all four datasets in an Australian benchmark study (bmchealthservres.biomedcentral.com). In a Statistics Canada census-to-mortality linkage the lowest linkage rate (77%) fell in ages 25 to 44 (www.statcan.gc.ca). Harron and colleagues call linkage errors inevitable given the imperfect and transient nature of identifiers (discovery.ucl.ac.uk), and link them to lower match rates in more vulnerable populations (discovery.ucl.ac.uk). Women's names are more likely to change over a lifespan (www.ncbi.nlm.nih.gov). These are linkage-stage and identifier results, not blocking-recall results, and the audit treats them as hypotheses about where to look rather than as expected values.

Table 3 stratifies the fixture before and after the source change.

Stratum	Baseline captured / adjudicated	After date-format change
Script: Latin	7 / 8	5 / 8

Stratum	Baseline captured / adjudicated	After date-format change
Script: Cyrillic in one source	1 / 1	1 / 1
Birth date present in both	7 / 8	5 / 8
Birth date missing in one source	1 / 1	1 / 1
Age under 40 (2026)	3 / 4	2 / 4
Age 40 to 59	3 / 3	2 / 3
Age 60 and over	2 / 2	2 / 2
All pairs	**8 / 9**	**6 / 9**

Several strata have one to four members, so single-pair strata prove nothing about rates; their purpose is to name which kinds of pair to sample more heavily. The Cyrillic pair survives only because both sources share an address, a dependency the rerun makes visible. Counts, not percentages, are reported in every cell, and none receives an interval because the fixture is a constructed set.

The rerun after a source-format change

Source B's ISO date format (year-month-day) changes to day/month/year with slashes in the 2026-10-15 export, and the year-extraction step now finds no four-digit year. Rule 1 stops producing candidates entirely. Rule 2 is unchanged. The candidate set shrinks from 11 to 9 pairs, and the reduction ratio rises from 0.89 to 0.91, while candidate recall falls from 8/9 to 6/9. The audit lists the **newly hidden pairs by identifier: A1-B1 and A9-B9**. Both had been captured only by Rule 1 because their addresses differ between sources. A dashboard watching candidate counts or reduction ratio would have read the change as an improvement.

The rerun is therefore a standing part of acceptance: repeat the comparison on the same adjudicated pairs after any change to a source format, a normalisation step or a blocking configuration, and diff the two candidate sets at the pair level.

Decision Rule and the Audit Worksheet

Decision rule

Thresholds belong to the steward and the population, and should be fixed before the test. Published studies use their own thresholds, which are not universal: the 2022 comparison used a pair-completeness target of 0.9 for every tuned method (arxiv.org), and the dense-blocking benchmark compares maximum pair quality only among blockers that reach 90% pair completeness (arxiv.org). These are protocol choices in those studies, not recommended acceptance values.

- **Accept:** candidate pairs are exposed and exact (or the sampling is declared); reference pairs meet the inclusion rules and are independent of the candidate set; overall recall, and recall in every stratum with enough members to judge, meet the pre-set floors; the rerun after the most recent source change shows no newly hidden pairs.

- **Revise candidate coverage:** one or more named true pairs, or a whole stratum, is missing. The fixture lands here: A7-B7 is hidden at baseline, the under-40 stratum shows 3 of 4, and the rerun hides two more pairs.
- **Hold with an explicit unknown:** the system does not expose candidate pairs, the labels derive from the system's own output, or the reference set cannot be shown to be independent. Record the unknown with its reason (not exposed, blocked, not applicable) rather than leaving the cell empty.

Worksheet

The reusable worksheet is a table with one row per audited population and run, plus a pair-level appendix. Copy the fields as column headers.

```
population,source_pair,eligible_frame,eligible_count,source_dates,candidate_export_available,candidate_count
,adjudicator,evidence_basis,inclusion_rule,label_independent_of_candidates,sampling_design,stratum,numerator
,denominator,candidate_recall,reduction_ratio,omitted_pairs,rerun_after_change,newly_hidden_pairs,decision,u
nknown_and_reason
```

Filled for the fixture's baseline row: eligible frame "cross-source A by B", eligible count 100, candidate export available "yes", candidate count 11, label independent "yes (document number)", sampling design "constructed set, no interval", numerator 8, denominator 9, candidate recall 0.889, reduction ratio 0.89, omitted pair "A7-B7", decision "revise candidate coverage". The unfilled equivalent for an opaque resolver reads: candidate export available "no", numerator and denominator "not computed", decision "hold", unknown "candidate pairs not exposed by the system".

Data Analysis and Evidence

The quantitative literature supports the audit's design but not any universal threshold. This section lists what each source measured and what it did not.

- **Papadakis et al. (2020).** The survey separates blocking, filtering and hybrid techniques (arxiv.org) and, by its own account, concentrates on candidate selection (arxiv.org). It supplies the definitions used here, and the caution that pair completeness is an optimistic recall estimate under a perfect matcher. It is a survey, not an experiment, and reports no single recall figure for the field.
- **Papadakis et al. (2022).** The comparison, presented at the 2023 IEEE International Conference on Data Engineering (www.computer.org) under the title "Benchmarking Filtering Techniques for Entity Resolution" in its latest preprint version, evaluated 14 filters and 4 baselines on 10 real-world datasets (arxiv.org). Its arXiv abstract reports the superiority of blocking workflows and string similarity joins (arxiv.org). Without per-dataset tuning, schema-based settings fell short of the recall target in half the cases (arxiv.org). Scope limits are explicit: it considers only methods that need no labelled training instances (arxiv.org), only textual entity profiles, serial processing, and its scalability test tuned each technique only on the smallest synthetic dataset. The right reading is method-specific: configuration affects recall and precision, and the findings are not a universal benchmark.

- **A default-parameter benchmark.** Wang and colleagues instead stick to default parameters to assess universality (arxiv.org). Comparing it with the 2022 study shows that "recall" in a blocking paper depends on the tuning protocol as much as on the method.
- **Washington State linkage study.** On 51,401 verified links, exact date-of-birth blocking compared 167,998 pairs against 2,642,062,801 for no blocking and retained 92.6% to 96.7% of links depending on group (doh.wa.gov). The ground truth was defined by a verified identifier, and the report covers one jurisdiction and one pair of record types.
- **Hong Kong romanisation study.** Recall of 68.8% for one non-standardised romanisation against above 95% for standardised schemes (ijpds.org); a small single-sample study of one script pair.
- **Tutorial-scale evidence.** A pyJedAI tutorial reports recall after each pipeline step on one dataset; those outputs are demonstrations of stepwise measurement, not performance claims (pyjedai.readthedocs.io).
- **Label provenance.** Zobel's pooled-judgment analysis is the strongest published argument that labels from system output overestimate recall (www.cs.cornell.edu), and Bohensky's review recommends routinely reporting characteristics of unmatched records and a measure of linkage quality (www.ncbi.nlm.nih.gov).

Taken together the sources agree on the shape of the failure, hidden pairs concentrated in records with inexact or missing key fields, and disagree on magnitudes because populations, labels and tuning differ. No source reviewed supports a claim that any particular blocking method is adequate for an arbitrary population.

Implications and Future Directions

- **For approval workflows.** The audit moves the conversation from matcher scores to candidate coverage. A steward can require three artifacts with every resolver release: the exported candidate set, the adjudicated reference with its inclusion rules, and a pair-level diff against the previous release. The first is a property of the system and the other two are the steward's responsibility, which is why the test applies equally to a commercial resolver, an open-source library or an in-house script.
- **For identity systems that keep unknowns visible.** The decision "hold with an explicit unknown" mirrors a practice that systems of record share: an unknown is declared with a reason, not left empty. The audit is a way for any reader to test that kind of claim, for any resolver, from the outside. Readers who want a matching library, MDM or data-quality tooling should use those categories; this test assumes one already exists.
- **For research.** Three open problems recur in the sources. First, sampling designs for recall when true pairs are rare and the frame must be independent of the candidate set remain active work, and the sampling estimators cited above are one line of it. Second, group-level blocking measures are new, and their adoption in practice is untested. Third, benchmark datasets are rarely built with hidden pairs in mind, a concern raised in the dense-blocking benchmark paper (arxiv.org). A worksheet like the one above could be adopted as a reporting convention, but whether it is cited or reused by others is untested, and no adoption claim is made.

- **Limits of this report.** The fixture is synthetic and tiny by design, the strata have one to four members, the thresholds are the reader's, and the tool table reflects documentation read in October 2026, not deployed behavior. Arithmetic was verified by script and by hand, but a real population will have a different mix of failure modes.

Frequently Asked Questions (FAQs)

What is blocking in entity resolution?

Blocking is the candidate-generation step that restricts comparison to pairs sharing a key, window or similarity neighbourhood, reducing the quadratic comparison space (users.cecs.anu.edu.au) (archive.nyu.edu). It is evaluated separately from matching.

How should blocking in record linkage be evaluated?

Compare the exported candidate pairs with independently adjudicated true pairs, compute candidate recall and reduction ratio with the eligible frame declared, stratify, and rerun after any change (users.cecs.anu.edu.au).

What is blocking recall, pairs completeness and reduction ratio?

Blocking recall and pairs completeness name the same quantity: true pairs in the candidate set over all true pairs (users.cecs.anu.edu.au). Reduction ratio measures the fraction of the comparison space removed and ignores candidate quality (users.cecs.anu.edu.au).

How are entity resolution results audited for missed matches?

Start before scoring: confirm the candidate export exists, build a reference that is independent of the system, list omitted pairs by identifier, and report the numerator and denominator for each stratum. If the system hides its candidates, record the test as unavailable.

Is entity resolution candidate generation always lossy?

Blocking is described as an inherently approximate procedure (users.cecs.anu.edu.au), and exhaustive comparison avoids loss only at quadratic cost (users.cecs.anu.edu.au). The audit therefore asks how lossy a configuration is on a stated population.

Conclusion

Blocking decides which pairs the rest of an entity resolution pipeline is allowed to find. The acceptance test proposed here is deliberately narrow: define the population, export the candidate pairs, compare them with independently adjudicated true pairs, report candidate recall and reduction ratio with their frames, stratify the omissions, rerun after every source change, and decide accept, revise or hold. When candidate pairs cannot be observed, the finding is an unknown with a stated reason, not a pass.

The published evidence supports the design and warns against shortcuts. Candidate-stage recall caps downstream recall (spiral.imperial.ac.uk), labels drawn from system output overstate it (www.cs.cornell.edu), and loss concentrates where key fields are inexact or missing (doh.wa.gov). The synthetic fixture shows the practical symptom in miniature: a configuration with 8 of 9 pairs captured fell to 6 of 9 after a date-format change while its reduction ratio improved. Downstream cluster questions such as pairwise F1 and bridge records come after this gate, not instead of it.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://recordlinkage.readthedocs.io/en/stable/guides/data_deduplication.html
2. https://arxiv.org/pdf/1905.06167
3. https://users.cecs.anu.edu.au/%7EPeter.Christen/publications/christen2011indexing.pdf
4. https://arxiv.org/pdf/2202.12521
5. https://doh.wa.gov/sites/default/files/2025-05/Hyperblock.pdf
6. https://ijpds.org/article/view/3036
7. https://www.cs.cornell.edu/courses/cs574/2002fa/assignments.org/zobel-sigir98.pdf
8. https://researchportalplus.anu.edu.au/en/publications/a-survey-of-indexing-techniques-for-scalable-record-linkage-and-d
9. https://www.amnesty.org/en/latest/research/2024/04/entity-resolution-in-indias-welfare-digitalization
10. https://tilores.io/content/entity-resolution-reflections-on-the-most-common-data-science-challenge
11. https://www.moderndata101.com/blogs/entity-resolution-at-scale-deduplication-strategies-for-knowledge-graph-construction
12. https://docs.dataforseo.com/v3/keywords_data/google_ads/search_volume/live
13. https://blockingpy.readthedocs.io/en/latest/user_guide/evaluation_metrics.html
14. https://recordlinkage.readthedocs.io/en/stable/ref-index.html
15. https://spiral.imperial.ac.uk/bitstreams/cddac2c2-301a-40e8-a85c-5d25e45c5a9a/download
16. https://archive.nyu.edu/bitstream/2451/27823/2/CeDER-PP-2007-15.pdf
17. https://arxiv.org/pdf/1407.3191
18. https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2012_draisbach_adaptive_2.pdf
19. https://www.cs.cmu.edu/afs/cs/project/reinforcement/web/talks-2001/01-03-19.mccallum.html
20. https://moj-analytical-services.github.io/splink/topic_guides/blocking/blocking_rules.html
21. https://recordlinkage.readthedocs.io/en/stable/ref-evaluation.html
22. https://moj-analytical-services.github.io/splink/api_docs/inference.html
23. https://moj-analytical-services.github.io/splink/api_docs/blocking_analysis.html
24. https://moj-analytical-services.github.io/splink/charts/cumulative_comparisons_to_be_scored_from_blocking_rules_chart.html
25. https://docs.dedupe.io/en/latest/API-documentation.html
26. https://anhaidgroup.github.io/py_entitymatching/v0.4.0/user_manual/api/blocking.html
27. https://anhaidgroup.github.io/py_entitymatching/v0.4.0/user_manual/api/debugging_blocker_output.html
28. https://pyjedai.readthedocs.io/en/latest/tutorials/CleanCleanER.html

29. https://docs.zingg.ai/latest/stepbystep/verifyblocking
30. https://www.nist.gov/node/761376
31. https://arxiv.org/pdf/1703.00617
32. https://arxiv.org/pdf/2404.14831
33. https://arxiv.org/pdf/2008.04443
34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3271236
35. https://arxiv.org/pdf/2504.04266
36. https://spiral.imperial.ac.uk/handle/10044/1/46235
37. https://asaip.psu.edu?p=677
38. https://math.unm.edu/%7Ejames/Agresti1998.pdf
39. https://arxiv.org/pdf/2409.16410
40. https://www.icao.int/sites/default/files/publications/DocSeries/9303_p3_cons_en.pdf
41. https://www.archives.gov/research/census/soundex
42. https://www.nber.org/papers/w24019.pdf
43. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0306-8
44. https://pmc.ncbi.nlm.nih.gov/articles/PMC5837697
45. https://bmchealthservres.biomedcentral.com/counter/pdf/10.1186/s12913-018-3495-x.pdf
46. https://www.statcan.gc.ca/en/conferences/symposium2016/program/14739-eng.pdf
47. https://discovery.ucl.ac.uk/id/eprint/10060709/1/Goldstein_Challenges%20in%20administrative%20data%20linkage%20for%20research_VoR.pdf
48. https://arxiv.org/abs/1905.06167
49. https://www.computer.org/csdl/proceedings-article/icde/2023/222700a653/1PByBaiXR7O
50. https://arxiv.org/abs/2202.12521

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.