Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Entity Resolution Evaluation Beyond Pairwise F1

> 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 · 2026-09-22 · entity resolution evaluation metrics · entity resolution · pairwise f1 · b-cubed · cluster evaluation · record linkage · variation of information · generalized merge distance · error analysis

Original article: https://reliablecontext.com/articles/entity-resolution-evaluation-beyond-pairwise-f1



In this article

1. Executive Summary
2. Introduction and Background
3. The Evaluation Object and Error Taxonomy
4. Metrics Beyond Pairwise F1
5. Worked Example: How Metrics Disagree
6. Labeling Design and Estimation
7. Data Analysis and Evidence
8. Implementation Guidance and Acceptance Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

Entity-resolution evaluation is a decision problem, not a search for one universally best score. A resolver turns records into predicted clusters, while a reference partition states which records belong to the same real entity. **Over-clustering** falsely joins distinct entities, and **under-clustering** splits one entity into several records or clusters. The same result can therefore be acceptable for one use and unsafe for another. Menestrina, Whang, and Garcia-Molina formalized this problem in 2010 and showed, within their experiments, that common measures can rank results differently (pure.kaist.ac.kr) (dblp.uni-trier.de). Their hotel-data sample produced a concrete reversal: basic merge distance favored one result, while pairwise F1 favored the other (vldb.org). That is a paper result, not proof that reversals occur in every data set.

The practical answer is a **metric portfolio tied to harm**. Pairwise precision measures the purity of predicted same-entity links, pairwise recall measures recovered true links, and pairwise F1 takes their harmonic mean (vldb.org). Because a cluster of size (n) contains (n(n-1)/2) pairs, pairwise measures give large clusters quadratic influence (www.cs.cmu.edu). **B-cubed** instead computes a score from each record's predicted and reference clusters and averages across records (aclanthology.org). Exact-cluster scores ask whether complete entities are recovered, while **variation of information (VI)** measures partition distance as (H(R)+H(S)-2I(R,S)) (vldb.org). Generalized merge distance (GMD) makes split and merge costs explicit, so it is useful when harms are asymmetric.

The labeling design must match the estimand. Pair sampling directly estimates link decisions but can miss rare, consequential entities and offers weak cluster-level diagnosis. Fully resolving probability-sampled entities supports pairwise, B-cubed, cluster, and error estimates from one black-box sample. Binette and colleagues recommend sampling records with replacement and resolving associated clusters (arxiv.org). A cluster's per-draw probability is then its size divided by the record count (arxiv.org), which requires corresponding weights. Their **400-cluster** PatentsView application is an application-specific illustration, not a universal sample-size rule (arxiv.org).

Approval should require more than a point estimate. The acceptance packet should contain the metric rationale, labeled-set provenance, inclusion probabilities, reviewer disagreement, uncertainty intervals, subgroup results, known coverage gaps, and a reversal procedure. Reliable Context is relevant as a system-of-record option because its observable contract includes one canonical record per real thing, cited and dated evidence, declared unknowns, and replayable past states. Those properties can preserve the evidence behind an approval, but they do not substitute for a representative reference set or defensible metrics.

Introduction and Background

Entity resolution, also called record linkage or disambiguation, maps records that refer to the same real-world entity into clusters. Evaluation begins only after separating four objects: the **record population**, the **predicted partition**, the **reference partition**, and the **decision boundary** that says which records and time period the approval covers. Black-box evaluation assesses the clustering output without relying on system architecture. This is the right viewpoint for a context engineer or data steward who must defend an output rather than inspect an algorithm.

The central risk is structural. A false merge contaminates a predicted cluster with records from another entity. A missed match leaves one true entity fragmented. The two errors can have unequal consequences. A false merge may propagate another organization's attributes into a canonical record, while a false split may suppress a complete history or double-count an entity. Pairwise precision and recall name these tendencies, but their aggregation unit is a record pair. Exact-cluster, record-weighted, information-theoretic, and operation-cost measures ask different questions.

The literature supports caution rather than metric absolutism. The foundational paper is a peer-reviewed 2010 article (pure.kaist.ac.kr), and its experiments found no conflicts when each synthetic series varied only one error mode (www.vldb.org). Conflicts appeared in particular cross-error comparisons, including **52 of 100** pairs in one synthetic comparison (www.vldb.org). The defensible conclusion is narrow: metric choice can change a ranking when candidate results contain different mixtures of errors.

This report develops an evaluation framework around that fact. It defines the objects, compares metrics by their weighting and failure sensitivity, calculates a worked example, connects labeling to the estimand, and ends with evidence required for acceptance.

The Evaluation Object and Error Taxonomy

Define the population before the score

A predicted clustering is meaningful only with a declared universe. The evaluator should specify included sources, record types, temporal cutoff, exclusions, and the treatment of records that cannot be reviewed. The official ER-Evaluation labeling guide requires a defined temporal or regional scope (er-evaluation.readthedocs.io). It also requires sampled reference clusters to include every relevant mention (er-evaluation.readthedocs.io). An incomplete reference cluster converts unlabeled records into apparent false positives or false negatives.

Four boundaries should be written before labeling:

- **Population:** Every record to which the acceptance claim will generalize.
- **Prediction:** The complete set of system clusters for that population, not only reviewed intersections. The estimator documentation explicitly requires the entire predicted disambiguation (er-evaluation.readthedocs.io).
- **Reference:** Human-resolved entities or adjudicated pairs with provenance and observation dates.
- **Decision:** The downstream action, its error costs, and the threshold that triggers approval, rejection, or limited use.

Separate lumping from splitting

Over-clustering, also called lumping, adds records that do not belong to the true entity. Under-clustering, also called splitting, omits records that do belong. In an inventor-disambiguation formulation, lumping is the addition of identifiers relative to the benchmark, while splitting is the absence of benchmark identifiers from the trial result (arxiv.org). These directions correspond approximately to precision and recall, but not every metric weights them equally.

Useful harm questions include:

- **False-merge harm:** What happens if two real entities share a canonical identifier?
- **False-split harm:** What happens if one entity remains in multiple clusters?
- **Cluster-size harm (hypothetical):** Does one error affecting a 500-record entity matter more than five errors affecting doubletons?
- **Singleton harm:** Must the evaluation credit correct isolation of records with no matches?
- **Reversal cost:** Can a decision be corrected, and can every downstream value be traced to the prior identity state?
- **Coverage harm:** Which regions, scripts, source systems, ages, or entity types may be missing from the reference sample?

The answers define the evaluand. A fraud-screening join, a research bibliography, and a canonical supplier register may rationally choose different weights, even when they use the same resolver. This report does not infer acceptability from an algorithm name or vendor benchmark.

Metrics Beyond Pairwise F1

No metric is neutral about its evaluation unit. *Table 1* compares six useful families. “Sensitivity to large clusters” describes the mathematical weighting, not an empirical guarantee for every data distribution.

Metric family	Unit weighted	Large-cluster sensitivity	False merge versus false split	Ground truth required	Interpretation
Pairwise precision, recall, F1	Same-cluster record pairs	High, because pairs grow quadratically with cluster size	Precision emphasizes false merges; recall emphasizes false splits; F1 balances them symmetrically	Pair labels or resolved clusters	Fraction of predicted links correct, fraction of true links recovered, and harmonic mean (vldb.org)
B-cubed precision, recall, F1	Individual records	Linear record weighting under equal record weights	Each record receives separate purity and completeness credit	Resolved clusters	Average record experience; singletons count (aclanthology.org)
Exact-cluster precision, recall, F1	Whole clusters	Small and large entities each count once	Any extra or missing record makes that cluster non-exact	Fully resolved clusters	Fraction of predicted entities exactly correct and fraction of true entities exactly recovered (vldb.org)
Homogeneity, completeness, V-measure	Entropy of class-cluster overlap	Depends on mass distribution, not raw cluster count alone	Homogeneity diagnoses mixing; completeness diagnoses fragmentation; beta can change emphasis	Full labels on evaluated records	Conditional-entropy purity, coverage, and their harmonic mean (scikit-learn.org)
Variation of information	Information mass in both partitions	Large clusters contribute through probability mass and logarithms	Symmetric distance decomposes into two conditional entropies	Full reference partition	Bits of information lost and gained between partitions (sites.stat.washington.edu)
Generalized merge distance	User-defined split and merge operations	Determined by cost functions	Costs can encode asymmetric harm directly	Full reference partition and declared costs	Minimum cost of a legal transformation from result to reference (www.vldb.org)

The table shows why a single leaderboard number cannot defend every decision. Pairwise F1 is informative when links are the action unit. Exact-cluster recall is demanding when a downstream process needs whole entities. B-cubed is often legible for record-level harm. VI supplies a proper distance, and GMD can price operations. A portfolio should keep directional components visible rather than report only a harmonic mean.

Pairwise measures

Let (Pairs(P)) be the unordered pairs placed together by partition (P). Pairwise precision is |(Pairs(R) \cap Pairs(S))| divided by |(Pairs(R))|, and pairwise recall uses |(Pairs(S))| as the denominator, where (R) is the prediction and (S) the reference (vldb.org). The F1 score is (2PR/(P+R)). It hides the direction of error, so report **precision and recall beside F1**.

Quadratic pair counts are both a feature and a limitation. They make a bad large merge visible, but a handful of large entities can dominate a population of small clusters. The Rand index adds correctly separated pairs, which can make abundant nonmatches influential. Adjusted Rand index corrects chance agreement, with random labeling near zero (scikit-learn.org). It also penalizes unnecessary pure splits (scikit-learn.org).

Record and entity-centric measures

B-cubed asks, for each record (i), how much its predicted cluster overlaps its true cluster. Per-record precision is |(R_i \cap S_i)|/|(R_i)| and recall is |(R_i \cap S_i)|/|(S_i)|; aggregate precision and recall average across records (aclanthology.org). Unlike link-only MUC, B-cubed credits singletons. A 2024 analysis also notes a nonzero floor because every record overlaps with itself (link.springer.com).

Exact-cluster scoring is stricter. A predicted cluster contributes only if the reference contains the identical set. Cluster precision divides exact matches by predicted cluster count, and cluster recall divides by reference cluster count. This makes the score directly intelligible for “complete entity” decisions but brittle for exploratory comparison.

CEAF, the Constrained Entity-Alignment F-measure, finds a one-to-one alignment between predicted and reference entities using maximum bipartite matching (aclanthology.org). Entities outside the optimum receive no credit (aclanthology.org). Mention-based and entity-based variants weight different objects, reinforcing the need to name the exact formula.

Information and operation distances

VI equals (H(R)+H(S)-2I(R,S)), or equivalently (H(R|S)+H(S|R)). It is nonnegative, symmetric, zero only for identical partitions, and satisfies the triangle inequality (sites.stat.washington.edu). For (n) records, its upper bound is \log (n), reached between the all-singleton and all-in-one partitions (sites.stat.washington.edu).

Basic merge distance counts binary splits and merges required to transform one partition into another, with splits preceding merges (vldb.org). GMD replaces unit counts with cost functions (f_m(x,y)) and (f_s(x,y)). Basic merge distance is the special case where every operation costs one (vldb.org). This is valuable when merging two large entities is more harmful than merging two singletons, or when split and merge reversal costs differ.

Worked Example: How Metrics Disagree

Consider eight records with reference entities (T_1={A,B,C}), (T_2={D,E}), (T_3={F,G}), and (T_4={H}). The predicted clusters are (P_1={A,B,D}), (P_2={C}), (P_3={E,F,G}), and (P_4={H}). The prediction makes two false merges and splits two true entities.

Table 2 exposes the overlap matrix from which all calculations follow.

Predicted cluster	Records	Overlap with T1	Overlap with T2	Overlap with T3	Overlap with T4
P1	A, B, D	2	1	0	0
P2	C	1	0	0	0
P3	E, F, G	0	1	2	0
P4	H	0	0	0	1

The matrix separates error direction. P1 mixes T1 with T2, P3 mixes T2 with T3, T1 is split across P1 and P2, and T2 is split across P1 and P3. P4 is an exact singleton. This one assignment therefore lets pair, record, cluster, information, and operation metrics emphasize different aspects.

Pairwise calculation

The prediction asserts six same-cluster pairs: AB, AD, BD, EF, EG, and FG. Only AB and FG are true, so **TP = 2** and **FP = 4**. The reference has five true pairs: AB, AC, BC, DE, and FG. Three are missed, so **FN = 3**.

- **Pairwise precision:** (2/(2+4)=0.333).
- **Pairwise recall:** (2/(2+3)=0.400).
- **Pairwise F1:** (2(0.333)(0.400)/(0.333+0.400)=0.364).

These definitions follow the set-of-pairs formulation in Menestrina et al. (vldb.org). The result looks poor because every three-record predicted cluster contributes three links and most are wrong.

B-cubed calculation

For A and B, precision and recall are both (2/3). D has precision (1/3) and recall (1/2). C has precision 1 and recall (1/3). E has precision (1/3) and recall (1/2). F and G each have precision (2/3) and recall 1. H has precision and recall 1.

- **B-cubed precision:** ((2/3+2/3+1/3+1+1/3+2/3+2/3+1)/8=0.667).
- **B-cubed recall:** ((2/3+2/3+1/2+1/3+1/2+1+1+1)/8=0.708).
- **B-cubed F1:** (2(0.667)(0.708)/(0.667+0.708)=0.687).

The **0.687** B-cubed F1 is far above the **0.364** pairwise F1 because every record gets one vote, including singleton H, rather than every predicted link getting one vote. B-cubed's aggregation by individual mention or record is documented in the coreference evaluation review (aclanthology.org).

Exact clusters, VI, and merge distance

Only {H} is exactly common to the four predicted and four true clusters. Exact-cluster precision, recall, and F1 are therefore each **1/4 = 0.250**. This score asks a stricter question than either pairwise or B-cubed scoring.

Using base-2 logarithms, (H(T)=1.906) bits, (H(P)=1.811) bits, and (I(T;P)=1.217) bits. Therefore **VI = 1.906 + 1.811 - 2(1.217) = 1.283 bits**. VI is a distance, so lower is better. Its conditional components are (H(T|P)=0.689) and (H(P|T)=0.594), which expose information in both error directions.

With unit-cost legal binary operations, split P1 into {A,B} and {D}, split P3 into {E} and {F,G}, merge {A,B} with {C}, then merge {D} with {E}. The basic merge distance is therefore **4 operations**. GMD could assign larger costs to the two false merges than to the two repairs. The requirement that a newly merged cluster fit inside a final reference cluster makes the transformation path legal (www.vldb.org).

The example does not identify a winner. It shows that scores answer distinct questions. An acceptance decision should state which question matters before seeing the numbers.

Labeling Design and Estimation

Pair sampling versus resolved-entity sampling

Pair sampling is economical when the estimand is pairwise precision or recall. Sampling predicted links estimates precision; sampling reference links estimates recall. However, random pairs from a sparse population are overwhelmingly nonmatches, and labeled pairs do not automatically reveal the complete extent of a cluster error. Application-specific enrichment can help find matches, but it changes inclusion probabilities and requires design weights.

Entity-centric sampling resolves whole entities. Binette et al. define the objective as a probability sample of known ground-truth clusters (arxiv.org). Reviewers identify records wrongly included in the predicted cluster and records wrongly omitted, then reconstruct the true entity. Fully resolved clusters imply positive labels for every pair inside them (arxiv.org). The same sample can support pairwise, B-cubed, cluster, and error estimates.

The recommended design samples records with replacement and resolves associated clusters. A true cluster of size |(c)| then has per-draw inclusion probability |(c)|/(N), so larger entities are selected more often. The implementation uses inverse cluster size weights for that probability-proportional-to-size design (er-evaluation.readthedocs.io). Uniform cluster sampling is a different estimand and must use different weights.

A labeling-plan worksheet

Table 3 is a pre-registration worksheet. Fill every cell before scoring. The final row records the operational evidence available from a system of record; it does not replace sampling or adjudication.

Field	Required entry	Why it matters
Population and cutoff	Record sources, exclusions, entity types, and observation date	Defines where the estimate may generalize
Sampling unit	Pair, record, predicted cluster, or reference entity	Determines the estimand and what can be diagnosed
Inclusion probability	Formula and value for every selected unit	Enables unbiased or design-consistent weighting
Number reviewed	Draws, unique units, records, positive pairs, and clusters	Makes effective sample size visible
Reference construction	Search procedure, stopping rule, and evidence retained	Exposes incomplete-cluster risk
Review disagreement	Independent review rate, raw agreement, adjudications, and unresolved cases	Separates label uncertainty from resolver error
Confidence interval	Estimator, resampling unit, confidence level, and finite-population treatment	Prevents false precision; NIST defines 95% coverage through repeated samples (www.itl.nist.gov)

Field	Required entry	Why it matters
Subgroups	Predeclared groups, sample counts, estimates, and interval widths	Finds concentrated harm without post-hoc selection
System-of-record evidence: Reliable Context	Canonical identity state, cited source, observation date, declared unknowns, and replayable prior state	Preserves what was approved and why; every value is described as exporting with chain of custody

The worksheet prevents a common category error: treating a metric implementation as an evaluation design. Software can compute a ratio correctly while the labeled sample targets the wrong population. A random-sample statistic estimates its corresponding population parameter only through the sampling design (www.itl.nist.gov).

Uncertainty and reviewer disagreement

Report a standard error or confidence interval for every acceptance metric. The ER-Evaluation functions return both a metric estimate and an estimated standard deviation (er-evaluation.readthedocs.io). Ratios of expectations require bias and variance adjustments under the sampled-cluster design (arxiv.org).

The interval method must follow the sampling unit. NIST's entity-linking evaluation bootstrapped documents for set metrics but did not extend those intervals to global clustering metrics when the independence rationale failed (tac.nist.gov). Normal approximations may also be inaccurate with very few failures or a very small sample (www.itl.nist.gov).

Reviewer evidence should include:

- **Independent overlap:** A stated fraction of units labeled by at least two reviewers.
- **Disagreement categories:** Same-entity disagreement, scope disagreement, insufficient evidence, and search incompleteness.
- **Adjudication:** A written rule and retained evidence for each resolved conflict.
- **Unresolved labels:** Excluded, interval-bounded, or analyzed as a sensitivity condition, never silently treated as negative.
- **Drift control:** Re-review after source, population, or policy changes.

Data Analysis and Evidence

The quantitative literature illustrates why evaluation design cannot be reduced to a threshold. Menestrina et al. generated a **10,000-entity** Zipfian reference with exponent **1.5** and ten progressively degraded results for each of three error types (www.vldb.org). Within single-error series, the measures moved monotonically. Across error types, the reported average conflict probability over their full synthetic grid was **9.5%** (www.vldb.org). Those values describe that experimental construction, not a base rate for production systems.

Their real-data check used random samples of **5,000 records** from shopping and hotel data (www.vldb.org). In the hotel comparison, basic merge distances were **427 versus 435**, while pairwise F1 values were **0.34 versus 0.61**, yielding opposite rankings (www.vldb.org). In the shopping sample, the authors reported no ranking conflict (www.vldb.org). Together, the positive and negative results argue for checking reversal risk rather than assuming it.

A simulation suggested samples of **400 or 800** were necessary for reasonable interval coverage in that application (arxiv.org). The authors expressly designed the simulation for PatentsView, so those sample sizes are not a benchmark for other populations (arxiv.org).

A 2026 Frontiers paper adds a truth-set-free diagnostic called Case Count Metric System (CCMS), classifying baseline clusters as unchanged, merged, partitioned, or overlapping (www.frontiersin.org). The authors present it as complementary to accuracy measures, not a replacement (www.frontiersin.org). In one synthetic threshold experiment, lowering a parameter from **0.67 to 0.57** increased merged cases from **0 to 179**, while precision fell from **0.7708 to 0.0221** (www.frontiersin.org). Raising it to **0.77** produced **105 partitioned cases**, while recall fell from **0.8495 to 0.4945** (www.frontiersin.org). These are illustrations of directional diagnosis, not transferable operating thresholds.

Implementation Guidance and Acceptance Evidence

Choose metrics from the defended decision

A practical selection sequence is:

1. **Name the harmful action.** State what a false merge and false split cause downstream.
2. **Choose the unit.** Select pairs for link decisions, records for per-record experience, entities for whole-cluster recovery, or operations for repair cost.
3. **Keep direction visible.** Report precision and recall, homogeneity and completeness, or split and merge components separately.
4. **Add a counter-metric.** Pairwise F1 plus B-cubed or exact-cluster recall exposes size-weighting changes.
5. **Predeclare thresholds.** Set acceptance limits before computing the held-out results.
6. **Estimate uncertainty.** Use the actual sampling design and resampling unit.
7. **Test subgroups.** Preserve cross-boundary links. Restricting both partitions to subgroup records can inflate estimates by hiding those links (arxiv.org).
8. **Test reversals.** Re-rank candidate outputs under plausible cost functions and metric families.

Homogeneity and completeness offer a useful directional cross-check. Homogeneity is perfect when predicted clusters contain records from only one reference class (scikit-learn.org), while completeness is perfect when every reference class stays in one cluster (scikit-learn.org). The V-measure beta parameter can weight

completeness more strongly (scikit-learn.org). A cost-weighted GMD is more explicit when the organization can quantify repair or harm.

Assemble the acceptance packet

An approver should receive these artifacts:

- **Decision statement:** Intended use, population, cutoff, and unacceptable consequences.
- **Metric rationale:** Formula, unit weighted, direction, threshold, and counter-metric.
- **Prediction provenance:** Resolver version or run identifier, input snapshot, and parameter policy, without requiring internal architecture.
- **Reference provenance:** Reviewer instructions, searches performed, cited evidence, observation dates, and stopping rules.
- **Sampling record:** Frame, randomization method, inclusion probabilities, replacements, exclusions, and weights.
- **Review record:** Reviewer count, overlap rate, disagreement categories, adjudications, and unresolved cases.
- **Results:** Point estimates, directional components, confidence intervals, denominators, and effective sample sizes.
- **Coverage:** Population segments absent or thinly represented, plus records marked not found, blocked, or not applicable.
- **Subgroups:** Predeclared cuts, sample sizes, uncertainty, and cross-boundary treatment.
- **Sensitivity:** Alternate metrics, plausible cost functions, and any ranking reversal.
- **Error audit:** High-impact false merges, false splits, error categories, and recurrence counts. The entity-centric paper recommends a two-pass audit, first developing categories and then applying a coding strategy (arxiv.org).
- **Reversal procedure:** Owner, trigger, correction path, affected exports, and historical-state preservation.

Reliable Context can be evaluated against observable guarantees at this evidence layer. Its public page states that related records are linked rather than folded together because merging real things falsifies both, and that past states are reproducible. Those are inspectable acceptance properties. They neither prove accuracy nor reveal internal resolution mechanisms. If a reader needs a vector database, retrieval-augmented generation framework, data catalog, master-data suite, customer relationship manager, or enrichment vendor, this system-of-record evidence model is not a substitute.

Implications and Future Directions

Three practices follow from the evidence. First, evaluation should become **decision-relative but reproducible**. The metric rationale may vary by harm profile, while the labeled frame, formulas, weights, and evidence remain fixed and reviewable. Second, evaluation sets should be treated as versioned knowledge assets. New records, source changes, identity corrections, and revised reviewer policies can change both the reference partition and the valid scope of an old score.

Third, uncertainty needs two layers. Sampling uncertainty describes how estimates would vary under repeated selection. Label uncertainty describes disagreement about the truth itself. A narrow confidence interval around inconsistent labels is not strong evidence. Future evaluation tooling should propagate both, report interval width by subgroup, and flag when acceptance changes under plausible adjudications.

Truth-set-free comparison also has a bounded role. Scalar partition distances can say how much two runs differ, but not always whether the difference is merging, splitting, or broader reorganization. The 2026 CCMS authors identify this diagnostic limitation for Rand, adjusted Rand, VI, and normalized mutual information (www.frontiersin.org). A run-to-run change analysis can prioritize review, but only labeled evidence estimates accuracy.

Finally, black-box evaluation supports procurement and governance neutrality. A common approval protocol can evaluate deterministic rules, probabilistic linkage, learned models, and human-assisted processes from their clustering outputs without making claims about their internals.

Frequently Asked Questions (FAQs)

How should entity resolution be evaluated?

Define the population and decision first, construct a probability-based reference sample, report directional error components, add at least one metric with a different weighting unit, quantify uncertainty, and audit subgroups. Do not begin with an F1 threshold and search for a sample that satisfies it.

Why is pairwise F1 insufficient?

Pairwise F1 compresses precision and recall and weights entities through their numbers of pairs. It can be appropriate for link-level decisions, but it hides direction and can let large clusters dominate. The 2010 experiments show that its rankings can conflict with merge-distance rankings in particular mixed-error comparisons (vldb.org).

Is B-cubed the best replacement?

No. B-cubed gives every record a local precision and recall view and counts singletons, making it useful when record experience is the target. Exact-cluster metrics are better aligned with whole-entity recovery, VI with partition distance, and GMD with explicit split and merge costs. Metric choice follows the defended decision.

How many entities need labeling?

There is no universal number. Sample size depends on prevalence, cluster-size distribution, target interval width, sampling design, weights, and subgroup requirements. The **400 or 800** figures in Binette et al. came from a PatentsView-specific simulation and should not be exported as a rule (arxiv.org).

What belongs in entity-resolution error analysis?

At minimum, count false merges and false splits, inspect their cluster-size and subgroup distributions, categorize causes through a coded audit, and retain examples with adjudicated evidence. Report how often reviewers disagreed and how unresolved cases affect the result.

Conclusion

Entity-resolution evaluation is defensible when its metric, sample, and acceptance rule trace directly to the decision being protected. Pairwise precision and recall are suitable for link-level questions, B-cubed for record-weighted experience, exact-cluster measures for complete entities, VI for information distance, and GMD for explicit repair or harm costs. None is universally best.

The worked example demonstrates why: the same eight-record result scored **0.364 pairwise F1**, **0.687 B-cubed F1**, **0.250 exact-cluster F1**, **1.283 bits of VI**, and **four** unit-cost merge-distance operations. Those values do not contradict one another. They describe different evaluation objects.

Approval should therefore depend on a predeclared portfolio, a probability-based labeled design, directional error reporting, uncertainty intervals, subgroup coverage, and preserved evidence. A result is not safe merely because one aggregate score is high. It is acceptable only when the evaluation shows that the relevant false merges, false splits, and unknowns fall within the harm limits of the stated use, and when the organization can reproduce what was approved and reverse it if the evidence changes.

Source links

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

1. https://pure.kaist.ac.kr/en/publications/evaluating-entity-resolution-results
2. https://dblp.uni-trier.de/rec/journals/pvldb/MenestrinaWG10.html
3. https://vldb.org/pvldb/vol3/R18.pdf
4. https://www.cs.cmu.edu/~hovy/papers/10BLANC-coref-metric.pdf
5. https://aclanthology.org/W10-4305.pdf
6. https://arxiv.org/html/2404.05622
7. https://er-evaluation.readthedocs.io/en/latest/06-data-labeling.html
8. https://er-evaluation.readthedocs.io/en/latest/03-estimating_performance.html
9. https://arxiv.org/pdf/1601.01963
10. https://scikit-learn.org/1.9/modules/generated/sklearn.metrics.homogeneity_completeness_v_measure.html
11. https://sites.stat.washington.edu/mmp/Papers/compare-jmva-revised.pdf

12. https://scikit-learn.org/1.4/modules/generated/sklearn.metrics.adjusted_rand_score.html
13. https://aclanthology.org/P98-1012.pdf
14. https://link.springer.com/article/10.1007/s10791-024-09436-7
15. https://aclanthology.org/anthology-files/pdf/H/H05/H05-1004.pdf
16. https://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm
17. https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_Trilingual_Entity_Discovery_and_Linking_overview.proceedings.pdf
18. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
19. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2026.1736939/full

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.