Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Data Fusion Error Propagation Benchmark: Upstream Error Audit

> 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 · 2026-10-20 · data fusion error propagation benchmark · data fusion benchmark · upstream data errors · error propagation in data pipelines · truth discovery · entity resolution benchmark · data provenance · oracle substitution · MaDI-Bench · data lineage

Original article: https://reliablecontext.com/articles/upstream-errors-data-fusion-benchmark-audit



In this article

1. Executive Summary
2. Introduction and Background
3. Defining Fused-Value Correctness and the Upstream Stages
4. What MaDI-Bench Labels and Scores
5. The Products Example as an Upstream Bound
6. An Oracle-Substitution Protocol for Fusion Errors
7. Data Analysis and Evidence
8. Limits of Public Benchmarks for Private Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

A low fused-value score is ambiguous. It can mean that a conflict resolver chose the wrong claim, or that the correct claim never reached the resolver because schema alignment, value normalization, blocking or entity matching removed it first. This report examines how the Mannheim Data Integration Benchmark (MaDI-Bench) of Steiner, Peeters and Bizer separates those cases, and proposes an oracle-substitution protocol that attributes **recoverable** final errors stage by stage. The protocol is a proposed analysis. No result in this report is a reproduced measurement of any real system.

MaDI-Bench supplies stage-level labels for five end-to-end tasks: a gold schema mapping, labeled normalization cells, labeled record pairs, and human-verified fusion records (www.uni-mannheim.de). Its first version, dated 29 June 2026 (blog.pebblous.ai), reported fusion accuracy between **40.2 and 84.9** (arxiv.org). That is the minimum and maximum of **15 cells** (three pipelines on five tasks), not a confidence interval, and the current version states that each pipeline is run once per task (wbsg-uni-mannheim.github.io). The revised version of 5 October 2026 (arxiv.org) has four pipelines (wbsg-uni-mannheim.github.io), and its range is **45.02 to 91.11** over 20 cells (arxiv.org). The definition of fusion accuracy also changed between versions, from an unweighted mean of per-attribute accuracy to a pooled share of reference cells (wbsg-uni-mannheim.github.io). Scores from the two versions should not be pooled.

The benchmark's own Products example shows the bound. A verified read speed of 130 sat in an unmatched record, so no conflict-resolution rule could recover it (arxiv.org). Observational pipeline comparisons suggest such losses matter, but they do not measure how many final errors each stage explains. Dong and colleagues had already separated factual source errors from extractor-introduced entity-linkage and schema errors, and found in a sample of **25** false triples that only **4%** were provided by the web sources (arxiv.org).

The recommended design reruns the same held-out entities with gold schema mapping, gold normalization and gold clusters supplied one stage at a time and in combination, then reports the **error-repair fraction** with an entity-level paired bootstrap. In a synthetic worked example (800 values, 213 baseline errors), single-stage

oracles repaired **12.7%** (schema), **48.4%** (normalization) and **15.0%** (clusters) of errors, while **21** errors were repaired only by combinations. Section "Implications and Future Directions" closes with a minimum reporting checklist for a defensible fusion claim.

Introduction and Background

Data fusion is the final step of data integration. After schema matching and duplicate detection, it combines multiple, possibly contradicting representations of one real-world object into a single consistent representation (hpi.de). Truth discovery, the closely related literature, is defined as finding the truth for each object by resolving conflicts among source values while estimating source reliability (arxiv.org). Its founding principle is that a web site is trustworthy if it provides many pieces of true information (research.ibm.com). Both literatures generally start from a table of claims already aligned to one schema and already grouped by entity. That starting point is the problem addressed here: in an end-to-end pipeline, the claims table is itself an output of earlier stages that make errors.

This report asks one question: when a fused value is wrong, can the evaluation tell whether the fault lies in conflict resolution or in an earlier stage? The question matters to researchers comparing knowledge-fusion or data-integration systems, because a low fused-value score is read as a statement about fusion methods. If a large share of the loss is upstream, a better resolver cannot recover it, and the comparison is partly a comparison of schema matchers, normalizers, blockers and entity matchers.

The report is an evaluation protocol and literature critique. It reconstructs what MaDI-Bench labels and scores, explains the Products missing-record example, sets out an oracle-substitution experiment with an error ledger and paired uncertainty, and discusses the limits of public benchmarks for private evidence. It also answers the practical questions that surround the topic: how to define fusion accuracy, how provenance and lineage support an audit of a fused value, and how entity resolution benchmark accuracy relates to downstream value accuracy.

Two scoping statements apply. First, MaDI-Bench is relational and its pipelines process tables, whereas the knowledge-fusion setting of Dong and colleagues processes triples from web extractors; the comparison below keeps the two layers apart. Second, the publisher of this report, Reliable Context, is a system of record for organizations that fuse private and public information. It is not an entrant in any benchmark discussed here, and no benchmark result is claimed for it. The protocol is written to be usable with any system.

Defining Fused-Value Correctness and the Upstream Stages

The final value and the stages that determine its candidates

A **fused value** is the single value a pipeline emits for one attribute of one real-world entity. It is **correct** when it matches a gold value under a stated comparator (exact, numeric tolerance, set equality). Four stages determine which claims can compete for that value:

- **Schema alignment:** maps source columns to target attributes. A wrong or missing mapping removes a source's claim, or places it under the wrong attribute.

- **Value normalization:** converts source values to the target format and taxonomy. Without it, a correct claim may differ textually from the gold value or split votes.
- **Blocking:** limits which record pairs are compared. Papadakis and colleagues define pair completeness as the share of detectable duplicates that survive, and note that it presumes a perfect matcher: a duplicate pair is detected as long as the records share at least one block (arxiv.org).
- **Entity matching and clustering:** decides which records describe one entity. A missed match leaves a record outside the cluster; a false match adds another entity's claims.

Only then does **conflict resolution** choose among the surviving claims, often by conflict resolution functions specified and applied per column (hpi.de). The MaDI-Bench authors state the same decomposition: a fused value can be wrong because matching kept the record with the correct value out of the cluster, because matching added records of other entities, or because conflict resolution chose a wrong value (wbsg-uni-mannheim.github.io).

Source error versus integration error

A fourth cause sits outside the pipeline. The source itself may state a wrong value. Dong, Gabrilovich and colleagues distinguish data fusion from knowledge fusion, and their abstract notes that extractors perform entity linkage and schema alignment, which adds noise unlike the factual errors that data fusion traditionally considers (arxiv.org). Provenance vocabularies give the formal handle: the W3C PROV data model defines a derivation as a transformation of an entity into another or the construction of a new one from a pre-existing one (www.w3.org), and each pipeline stage is an activity that acts upon or with entities (www.w3.org).

An evaluation that separates the causes needs a label for each stage, not only a final gold value. *Table 1* below maps each stage to its gold artifact and metric in MaDI-Bench and notes the sample denominators that matter for uncertainty.

Stage	Input to output	Gold artifact in MaDI-Bench	Metric and denominator
Schema matching	Source columns to target attributes	One gold schema mapping per task	F1 over column correspondences (wbsg-uni-mannheim.github.io)
Value normalization	Source value to target format	Labeled normalization cells and target taxonomies; 1,745 cells over five base tasks (wbsg-uni-mannheim.github.io)	Accuracy over test cells
Blocking	Records to candidate pairs	Labeled record pairs; more than 93,000 pairs over five tasks (wbsg-uni-mannheim.github.io)	Pair completeness and reduction ratio; blockers are selected on the validation set at 97% pair completeness (wbsg-uni-mannheim.github.io)
Entity matching	Candidate pairs to clusters	The same labeled pairs; test sets verified by hand except for Papers (matches derived from DOI identifiers) and Products (which reuses the WDC Products mapping)	F1 on the test pairs

Stage	Input to output	Gold artifact in MaDI-Bench	Metric and denominator
Data fusion	Clusters to one value per attribute	Human-verified fused records: 100 validation and 100 test entities per task (wbsg-uni-mannheim.github.io)	Fusion accuracy over verified cells

Table 1 shows why per-stage metrics are not enough on their own. Each metric is computed on its own sample, with its own denominator: thousands of labeled pairs for matching, but 100 test entities per task for fusion. A matcher's pair-level F1 does not state whether the particular record that carried a crucial value was linked to the right cluster. The gap between the two samples is the reason the protocol below holds the fusion entities fixed and intervenes on stages, rather than comparing stage scores across different samples.

What MaDI-Bench Labels and Scores

Construction and provenance of the gold labels

MaDI-Bench is described by its authors' university as the first benchmark for end-to-end integration of relational tables covering all steps of the integration process (www.uni-mannheim.de). Its fusion gold is human-verified, and the ground-truth reference covers only the test entities but gives the stronger guarantee of correctness (wbsg-uni-mannheim.github.io). The current fusion gold, described on the project site as version 2, leaves out the values that cannot be obtained from any source record of the entity, even after normalization (wbsg-uni-mannheim.github.io). This is a useful property for the present purpose, because it restricts the scored set to values that are in principle **recoverable**. It also means the benchmark scores no value for which every source is wrong, so it understates one cause of real-world failure.

The variant generator reinforces this design. For each fusion value it enforces a closeness floor, so that at least one surviving source value stays within an attribute-specific tolerance, and the benchmark offers an easy, a medium and a hard variant of each base task, 20 tasks in total.

Two versions, two denominators

The paper exists in two arXiv versions with different numbers, and the figure quoted in secondary commentary comes from the first. Version 1 reported fusion accuracy of 40.2 to 84.9. An independent September 2026 report cites that range as between 40.20 and 84.87% (blog.pebblous.ai), which matches version 1 only. The revised version has four pipelines and reports a range of 45.02 to 91.11 (arxiv.org). *Table 2* compares the two.

Property	Version 1 (29 June 2026)	Version 2 (5 October 2026)
Pipelines	Three: human (P1), best-of-breed (P2), LLM-based (P3)	Four: P1, P2, P3 and a coding-agent pipeline (P4)
Cells behind the range	15 (3 pipelines x 5 tasks)	20 (4 pipelines x 5 tasks)
Fusion accuracy range	40.20 to 84.87, over 15 cells	45.02 to 91.11, over 20 cells

Property	Version 1 (29 June 2026)	Version 2 (5 October 2026)
Fusion accuracy definition	Unweighted mean of per-attribute value accuracy (arxiv.org)	Share of reference cells whose fused value is correct, pooled over all attributes; a missing value counts as wrong (wbsg-uni-mannheim.github.io)
Runs	No repeated-run statement found	One per pipeline and task
Products, P1 fusion accuracy	40.20	71.12 (arxiv.org)

The two ranges are spreads over tested pipeline-by-task cells. They are not confidence intervals, and the project site states that each pipeline is run once per task (wbsg-uni-mannheim.github.io), so no run-to-run variance is reported. The Products entry for the human pipeline moved from 40.20 to 71.12 between versions. The paper documents changes in pipelines, the fusion gold and the metric definition, but a reader cannot decompose that shift from the published text. It is therefore a direct illustration of why a score is a property of a label set, a metric definition and a pipeline version together.

Per-task variation and other stage ranges

Variation across tasks and stages is large. The project site reports pair completeness between 93.33 and 100 for all pipelines and tasks (wbsg-uni-mannheim.github.io), and blockers are chosen for the highest reduction ratio among those that reach 97% pair completeness on the validation set. Reported blocking recall is therefore partly a tuning outcome. The version 1 figures for blocking, 94.37 to 100.00% across five tasks and three pipelines, are cited by the independent Pebblous report (blog.pebblous.ai).

Fully correct entities are much rarer than correct cells. The project site reports ground-truth fully-correct rates, which require every evaluated attribute of an entity to be right at once, between 0.10 and 0.25 (wbsg-uni-mannheim.github.io). It also notes that the silver and ground-truth references rank two pipelines differently on fusion accuracy (wbsg-uni-mannheim.github.io), so the choice of reference changes conclusions. The silver reference is the clusters of the human-engineered pipeline, which is a pipeline output and not a gold label.

The Products Example as an Upstream Bound

What the example shows

The paper uses its Products task to illustrate how matching errors propagate into fusion. Its example is a flash drive that appears in all four sources, where the matcher links only the first source's record, with a read speed of 100 MB/s, while the verified value of 130 sits in an unmatched record. Version 1 states the consequence directly: the missing record never enters the cluster, so no conflict-resolution rule can recover the correct value (arxiv.org).

The surrounding evidence is observational. Version 1 notes that best-of-breed led Products entity matching and that the two pipelines trailing it also trail in fusion, at 40.20% and 43.65% against 56.68% (arxiv.org). That is a rank association across three pipelines on one task. It supports the bound as a mechanism, and the

example demonstrates it for one value. It does not state what share of the Products fusion errors in any pipeline are explained by matching, nor whether normalization or schema errors contribute. Attributing that share is what the protocol below is for.

The same bound in blocking and in earlier fusion work

The upstream bound is well established at the blocking stage. An independent September 2026 report states that links blocking never generates are not matcher errors, and that no improvement in matching accuracy brings them back (blog.pebblous.ai). Its own headline result comes from a different, single-author benchmark preprint and concerns a hard stratum of links that its blocking schemes recovered poorly. That is the report's result on that benchmark and not a general blocking rate, so it is not reproduced here. As noted above, the pair-completeness metric presumes a perfect matcher, so it is an optimistic estimate of how many duplicates a later stage can still find.

Classical truth-discovery evaluations had the converse limitation: they began after alignment. On the Deep Web data, state-of-the-art fusion found correct values for 96% of data items on average (arxiv.org), but those experiments used sources that were already aligned and gold standards built from trusted sources, which the authors acknowledge can put wrong or coarse-grained values in the gold standard. An accuracy of 96% on aligned claims says little about a pipeline that must create the alignment.

Dong et al.: a different layer

Dong and colleagues studied triples from multiple extractors rather than relational tables. In a random sample of 25 false triples, only 4% were indeed provided by the web sources (arxiv.org); the rest were attributed to extraction errors in triple identification, entity linkage and predicate linkage. These results are web-scale and extractor-specific. They establish the conceptual need to attribute errors at separate layers. They do not validate MaDI-Bench's tasks, and the 25-triple sample is too small to transfer as a rate. The same authors found that their gold standard, built under a local closed-world assumption, marked some correct decisions wrong, a reminder that gold-label uncertainty is a cause in its own right.

Entity resolution benchmark accuracy as a proxy

Pair-level accuracy on entity resolution benchmarks is an unreliable proxy for downstream value accuracy. Köpcke, Thor and Rahm held blocking fixed for all systems (vldb.org), so their F-measures are conditional on a fixed candidate set, and built the product ground truth from entities with a valid UPC (Universal Product Code) (vldb.org). Even so, no approach exceeded an F-measure of 62% on the Amazon-GoogleProducts task (vldb.org), while bibliographic tasks were much easier. Their abstract concludes that product entities from online shops were not sufficiently solved by the approaches tested (vldb.org). A high matcher F1 on one dataset therefore says nothing about whether the record carrying a crucial value reached the correct cluster in another.

An Oracle-Substitution Protocol for Fusion Errors

Principle

An oracle substitution replaces one stage's output with its gold label while leaving everything else unchanged, then reruns the downstream stages. The idea is not new in pipeline evaluation. Finkel, Manning and Ng describe a greedy pipeline as one that causes errors to propagate (aclanthology.org), and the cited gold-versus-automatic comparison in semantic role labeling dropped from 91.2 to 80.0 when gold parses were replaced by automatic ones (aclanthology.org). A 2026 workshop paper proposes "oracle substitution" as a necessity test that substitutes in a corrected version of a stage input (openreview.net), and argues that walking each failed case through the pipeline and blaming the first stage fails to identify the stage whose correction removes most failures (openreview.net). These are a 2006 comparison and a 2026 workshop poster, and the protocol below is a proposed adaptation to data fusion rather than an established method.

Fixed set, permitted interventions, unchanged evidence

The design has five rules.

- **Fix the entity set.** Use the same held-out entities, with human-verified values, for every run. In MaDI-Bench that is the 100 test entities per task.
- **Keep source evidence unchanged.** Oracles may change mappings, normalized forms and cluster membership. They must never edit a source claim, or the incorrect-source-claim category disappears.
- **Permit three interventions:** the gold schema mapping (S), gold normalized values for the claims (N), and gold entity clusters (C). Blocking and matching are merged into C, because a gold cluster supersedes both.
- **Run the full factorial.** Three binary factors give 2^3 = 8 runs: none, S, N, C, S+N, S+C, N+C, S+N+C. A full factorial design lets the analyst estimate all main and interaction effects (www.itl.nist.gov), but the run count doubles with each added factor and the handbook does not recommend full factorials for five or more factors (www.itl.nist.gov). Splitting blocking from matching or adding variants of the resolver should therefore use a fractional design.
- **Hold the resolver fixed.** The conflict-resolution rule stays unchanged across runs, so that differences are attributable to its inputs.

The error ledger

For each baseline wrong value, record one ledger entry with the entity, attribute, source quote, observation date, selected value, retained alternatives and the verdict. The proposed categories are:

- **Absent candidate:** the gold value appears in no retained claim at baseline.
- **Wrong cluster:** the record carrying the gold value is outside the entity's cluster, or another entity's record is inside it.
- **Wrong normalized value:** a correct claim exists but differs from the gold form.
- **Schema error:** the claim is under the wrong attribute or unmapped.
- **Incorrect source claim:** no claim in the fully corrected candidate set matches the gold value.

- **Wrong conflict choice:** a correct claim is in the fully corrected candidate set but the resolver chose another.
- **Gold-value uncertainty:** annotators disagreed, or the gold value is coarse or contested.

The decisive rule is that the last three categories are assigned from the full-oracle run (S+N+C), where upstream causes have been removed by construction. This is why the benchmark's recoverable-value restriction matters: a value for which every source is wrong is an incorrect-source-claim error under any oracle.

Estimands and uncertainty

Let E0 be the set of final errors in the baseline run. For intervention set A, the **error-repair fraction** is the number of values in E0 that are correct under A divided by |E0| on the same fixed set. Two companion counts prevent over-reading: **regressions**, values correct at baseline that become wrong under A, and the **net change**. Regressions occur because a stage's errors can offset one another, so correcting one stage alone can expose a latent fault in another.

Values within an entity share a cluster and a source record, so they are not independent. A bootstrap that resamples entities, not values, respects that structure. The bootstrap needs only the observed paired outcomes and can be applied to any metric (aclanthology.org), and goes back to the sampling-distribution formulation of Efron (sites.stat.washington.edu). Sampling-based tests are generally preferred over parametric ones when distributional assumptions are violated (aclanthology.org). McNemar's test applies to paired binary outcomes (aclanthology.org) and, in Dietterich's comparison, is the only test with acceptable type I error for algorithms that can be executed only once (pubmed.ncbi.nlm.nih.gov), but it treats values as independent and its asymptotic form needs at least 10 discordant pairs (www.ncbi.nlm.nih.gov); it also compares paired marginal proportions and should not be used as a measure of agreement (www.ncbi.nlm.nih.gov). For proportions on a small fixed set such as 100 entities, Brown, Cai and DasGupta recommend the Wilson interval or the equal-tailed Jeffreys prior interval over the Wald interval (repository.upenn.edu), because the chaotic coverage properties of the Wald interval are far more persistent than is appreciated (repository.upenn.edu).

Interactions are exposed in two ways. The first is overlap: how many errors in E0 are repaired by two or more single interventions. The second is complementarity: how many errors are repaired by a combination but by none of its members. Attribution indices such as the Shapley value enumerate subsets of factors, which is exponential in the number of stages, and for three stages the eight-cell table is already a complete description. Factorial analysis treats interactions as first-class results, and the NIST handbook notes that besides the main effects it is important to check the interactions (www.itl.nist.gov). Subtracting stage scores from different samples hides interactions that the full table shows.

Data Analysis and Evidence

What the published evidence quantifies

The published numbers in this area quantify three different things, and they should not be mixed. Fusion accuracy on aligned claims is high in classical truth-discovery work (**96%** on the Deep Web data). Fusion accuracy on end-to-end MaDI-Bench tasks ranges from **45.02 to 91.11** across 20 pipeline-task cells in version

2. Error attribution at the extractor layer rests on a 25-triple sample. None of these decomposes final errors by stage on a common entity set. Related work on cascades is survey-based: a study of **53** practitioners reported data cascades with **92%** prevalence (research.google), a statement about reported experience and not a pipeline error rate. The same study found that 45.3% of practitioners reported two or more cascades in a given project (storage.googleapis.com).

Label quality also bounds what any score can show. The book-author gold standard of the source-dependence study covered 100 randomly selected books (lunadong.com), and the TruthFinder evaluation used the authors on the book cover as the standard fact (hanj.cs.illinois.edu). Chance-corrected agreement, defined as the proportion of the possible agreement beyond chance that was actually observed (aclanthology.org), is the standard way to report annotator reliability for gold values.

Worked example (Hypothetical Example)

This is a **Hypothetical Example**. The numbers below come from a synthetic simulation constructed for this report, with a fixed random seed. The simulated world has 100 entities, 8 attributes (800 gold values) and 4 sources. Each source claim is wrong with probability 0.30, non-normalized with probability 0.20, and each source-attribute mapping is wrong with probability 0.10. Each source record is missing from its entity's cluster with probability 0.10, and a foreign claim intrudes with probability 0.04. The resolver takes a plurality vote with a fixed source-priority tie-break. The parameters were chosen to illustrate the analysis. They are not estimates of any real system, and nothing here is a reproduced result.

At baseline, **587 of 800** values are correct (**73.4%**) and **213** are wrong. *Table 3* gives the eight-run intervention matrix. Intervals are 95% percentile intervals from 2,000 entity-level bootstrap resamples of the same fixed entity set.

Oracle supplied	Errors repaired	Regressions	Repair fraction (95% interval)	Final accuracy
None (baseline)	0	0	0%	73.4%
Schema (S)	27	13	12.7% (8.8 to 17.0)	75.1%
Normalization (N)	103	0	48.4% (41.9 to 55.1)	86.3%
Clusters (C)	32	8	15.0% (9.9 to 20.9)	76.4%
S + N	122	7	57.3% (51.2 to 63.5)	87.7%
S + C	64	18	30.0% (23.5 to 37.0)	79.1%
N + C	131	3	61.5% (54.8 to 67.9)	89.4%
S + N + C	148	10	69.5% (63.5 to 75.1)	90.6%

Table 3 shows three patterns that a subtraction of stage scores would miss. First, single-stage repairs do not add. The single interventions repair 27, 103 and 32 errors, a sum of 162, but their union is **129** distinct errors because 17 are repaired by both S and N, 4 by S and C, 14 by N and C, and 2 by all three. Second, the full oracle repairs **148** errors, of which 20 are repaired by no single oracle, so they need at least two stages corrected together. In total, **21** errors are repaired only by some combination: seven need exactly N and C,

seven need exactly S and N, four need S and C, and three follow other patterns; one of the 21 is not repaired under the full oracle. To reconcile the counts: the union of single-oracle repairs is 129, of which 1 is not repaired under the full oracle, leaving 128; adding the 20 combination-only errors that the full oracle does repair gives 148. Third, regressions are not zero. Supplying gold schema alone repairs 27 errors but breaks 13 values that were right, for a net gain of 14, because correcting one stage can expose another stage's fault or change which claim wins a vote.

After the full oracle, **65** errors remain. Applying the ledger rule, **9** are incorrect-source-claim errors (no correct claim exists among the corrected candidates) and 56 are wrong conflict choices (a correct claim is present but not selected), so in this synthetic world the conflict resolver accounts for at most **56 of 213** baseline errors, or about 26%. Two further baseline errors were repaired by a partial oracle but not under the full oracle, which shows that repair is not monotone in the number of stages corrected. An analyst who read the baseline score as a statement about conflict resolution would have attributed all 213 errors to it.

Limits of Public Benchmarks for Private Evidence

Provenance and lineage for fused values

An audit of a fused value needs a record of how it was derived. Provenance is information about the entities, activities and people involved in producing a piece of data (www.w3.org), and in databases why-provenance describes the source tuples that witness the existence of an output tuple (homepages.inf.ed.ac.uk), and PROV-O expresses that model in OWL2 so that provenance can be represented and interchanged across systems (www.w3.org). In the formulation of Buneman, Khanna and Tan, why-provenance gives the source data that had some influence on the existence of the output (link.springer.com), and where-provenance the location in the source from which a value was extracted (link.springer.com). How-provenance is more general than why-provenance (homepages.inf.ed.ac.uk), and semiring annotations propose polynomials to capture provenance (www.cs.ucdavis.edu). The survey of Cheney, Chiticariu and Tan describes provenance as describing the origins and history of data (homepages.inf.ed.ac.uk), also called lineage. For fusion evaluation the practical requirement is simple: every final value must retain the source quote, the observation date, the retained alternatives and the stages that touched it, so that the error ledger above can be built after the fact. Herschel and colleagues organize the field by what provenance is used for, including accountability and process debugging (link.springer.com).

Private evidence, shifting sources and unknowns

Public benchmarks cannot replace this ledger for private evidence, for four reasons.

- **Different noise.** Public benchmark sources can be partly known to model pretraining, a limitation the MaDI-Bench authors state (arxiv.org); private evidence is not, and its error profile is unmeasured.
- **Gold values come from the public web.** MaDI-Bench verifiers used web research outside the task sources. For a private attribute no outside source exists, so gold values need an internal adjudication process, with reported agreement.
- **Sources shift.** A gold value valid at one observation date can be stale later. The ledger needs an observation date per claim, and a rerun against a changed source is a different experiment.

- **Unknowns are real outcomes.** MaDI-Bench counts a missing value as wrong. In private settings, a value that is truly unknown, blocked or not applicable should be a distinct ledger outcome, not an empty cell scored as an error.

Copying between sources adds a further limit. Dong, Berti-Equille and Srivastava observed that a false value can be spread through copying (lunadong.com). The independence assumption behind many truth-discovery methods, that false information from different sources tends to differ, does not survive copying. Gold standards also differ in how they score: the book-author precision of that work was not comparable with TruthFinder's, which scored partial credit (lunadong.com). A reported fusion accuracy is therefore tied to its gold-standard construction rule, as the MaDI-Bench definition change between versions also shows.

Finally, stage rankings are dataset-dependent. MaDI-Bench itself reports that the best pipeline differs by task, and the independent evidence on entity matching above shows the same sensitivity to dataset construction. That is another reason to run the oracle protocol on the target data.

Implications and Future Directions

For researchers, the immediate implication is that a fused-value score should be reported with its upstream conditions. Two pipelines with equal fusion accuracy may differ sharply in how many errors were recoverable, and two with different accuracy may differ only in their matcher. Releasing stage-specific gold labels, as MaDI-Bench does, makes oracle runs possible without new annotation, because the gold schema mapping, labeled pairs and human-verified fusion records already exist. What is missing is record-level cluster gold for the fusion entities, so that C can be supplied as a true oracle and not as a matcher run on labeled pairs. Releasing that artifact would let the protocol be run reproducibly.

For practitioners combining private and public data, the same protocol can be run on a small internal gold set, with the observation dates and quotes retained. Publisher note: Reliable Context is a system of record for this setting. The protocol in this report does not depend on any particular system of record, and the publisher claims no benchmark result.

Minimum reporting checklist for a defensible fusion claim

- **Name the version.** State the benchmark version, date and pipeline versions, and do not pool scores across versions.
- **State the metric definition.** Say whether accuracy is per-attribute-mean or pooled, and how missing values and ties are scored.
- **State denominators.** Report entities, attributes and verified values for the fusion set, and the sample sizes behind each stage metric.
- **Report the spread honestly.** Distinguish a min-max over cells from a confidence interval, and report run-to-run variance if there is any.
- **Use a fixed entity set.** Compare ordinary and oracle runs on the same held-out entities.
- **Report all eight cells.** Include single and combined interventions, with changed-value counts, regressions and net change.

- **Use an entity-level paired bootstrap.** Report intervals for repair fractions, and state the number of resamples.
- **Show overlap and complementarity.** List errors repaired by more than one single intervention, and errors repaired only jointly.
- **Publish the ledger.** Each entry carries the source quote, the observation date, the selected value and the retained alternatives.
- **Separate source error from integration error.** Assign incorrect-source-claim and wrong-conflict-choice only from the full-oracle run.
- **Report gold uncertainty.** State annotator agreement, and mark values whose gold is coarse or contested.
- **Label the generator.** If part of the analysis is synthetic, say so, and do not present it as a measurement.

Future work includes extending the factorial design to split blocking from matching, with a fractional design as stages multiply; testing whether the interaction structure is stable across the easy, medium and hard variants that MaDI-Bench provides; and replicating the protocol on triple-based knowledge fusion, where the extractor layer replaces the relational pipeline.

Frequently Asked Questions (FAQs)

What is a data fusion benchmark, and which one covers the full pipeline?

A data fusion benchmark supplies gold values against which conflict resolution is scored. Most evaluate fusion in isolation on aligned claims. MaDI-Bench is described by its authors' university as the first covering all steps of relational integration (www.uni-mannheim.de).

How is data fusion accuracy evaluated?

Version 1 of MaDI-Bench defines it as an unweighted mean of per-attribute value accuracy, and the project site defines the current version as the share of the verified cells of the fusion test set for which the fused output holds a correct value. Both definitions need a human-verified reference, which is why gold-label construction belongs in the report.

Where do upstream data errors in fused data come from?

They come from the stages before conflict resolution: schema alignment, normalization, blocking and matching, plus the source claims themselves. The benchmark's Products example shows a verified value in an unmatched record that no resolver could recover.

How do truth discovery methods handle conflicting data sources?

They estimate source reliability while resolving conflicts. In the Bleiholder and Naumann line of work, conflict resolution functions are specified and applied per column (hpi.de). TruthFinder states the principle as a web site being trustworthy if it provides many pieces of true information (research.ibm.com).

How do provenance and lineage support auditability of a fused value?

Provenance describes the origins and history of data (homepages.inf.ed.ac.uk), and why-provenance names the source tuples that witness the existence of an output tuple (homepages.inf.ed.ac.uk). A fused value that retains these links lets a reviewer rebuild the error ledger.

How accurate are entity resolution benchmarks as a guide to downstream values?

Only loosely: no approach exceeded an F-measure of 62% on one product-matching task in a classic evaluation (vldb.org), and pair-level scores do not state which records carried which values.

How should private and public data quality be assessed when merged?

Use a small internally adjudicated gold set, keep a source quote and observation date for each value, and run the oracle protocol. Public benchmark scores do not transfer, for the reasons given above.

Conclusion

A fused-value score is a joint outcome of alignment, normalization, clustering and conflict resolution, and a published number does not say which stage contributed what. MaDI-Bench provides stage labels and a first end-to-end view, and its Products example demonstrates the upstream bound for one value. Its reported range is a min-max over tested cells, differs between versions, and rests on single runs. Earlier knowledge-fusion work established the principle of attributing errors at separate layers but on a different data model and scale.

The proposed oracle-substitution protocol closes the gap by holding entities and source evidence fixed, supplying gold schema, normalization and clusters in all eight combinations, and reporting repair fractions with entity-level paired intervals, overlap, complementarity and regressions. In the synthetic example, conflict resolution accounted for about a quarter of the baseline errors, and 21 errors could be repaired only by joint corrections. Those numbers illustrate the analysis and are not findings about any system. Researchers who adopt the checklist can state whether a low fused-value score reflects conflict resolution itself or errors that arrived before it.

Source links

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

1. https://www.uni-mannheim.de/dws/news/mannheim-data-integration-benchmark-madi-bench-released
2. https://blog.pebblous.ai/report/corporate-family-resolution-blocking-gap/en
3. https://arxiv.org/html/2606.30371v1
4. https://wbsg-uni-mannheim.github.io/MaDI-Bench
5. https://arxiv.org/abs/2606.30371
6. https://arxiv.org/html/2606.30371
7. https://arxiv.org/pdf/1503.00302
8. https://hpi.de/naumann/projects/completed-projects/fusem.html
9. https://arxiv.org/pdf/1505.02463

10. https://research.ibm.com/publications/truth-discovery-with-multiple-conflicting-information-providers-on-the-web
11. https://arxiv.org/pdf/1905.06167
12. https://arxiv.org/abs/1503.00302
13. https://www.w3.org/TR/prov-dm
14. https://arxiv.org/pdf/1503.00303
15. https://vldb.org/pvldb/vol3/E04.pdf
16. https://aclanthology.org/W06-1673.pdf
17. https://openreview.net/forum?id=hvnTPAlx4T
18. https://www.itl.nist.gov/div898/handbook/pri/section3/pri3332.htm
19. https://www.itl.nist.gov/div898/handbook/pri/section3/pri333.htm
20. https://aclanthology.org/D12-1091.pdf
21. https://sites.stat.washington.edu/courses/stat527/s13/readings/ann_stat1979.pdf
22. https://aclanthology.org/P18-1128.pdf
23. https://pubmed.ncbi.nlm.nih.gov/9744903
24. https://www.ncbi.nlm.nih.gov/sites/books/NBK560699
25. https://repository.upenn.edu/handle/20.500.14332/47517
26. https://www.itl.nist.gov/div898/handbook/pri/section6/pri633.htm
27. https://research.google/pubs/everyone-wants-to-do-the-model-work-not-the-data-work-data-cascades-in-high-stakes-ai
28. https://storage.googleapis.com/gweb-research2023-media/pubtools/5936.pdf
29. https://lunadong.com/publication/dependence_vldb.pdf
30. https://hanj.cs.illinois.edu/pdf/kdd07_xyin.pdf
31. https://aclanthology.org/J08-4004.pdf
32. https://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf
33. https://www.w3.org/TR/prov-o
34. https://link.springer.com/chapter/10.1007/3-540-44503-X_20
35. https://www.cs.ucdavis.edu/%7Egreen/papers/pods07.pdf
36. https://link.springer.com/article/10.1007/s00778-017-0486-1

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.