Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Truth Discovery with Copying Sources: Benchmark Guide

> 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 · truth discovery benchmark · truth discovery algorithms · source dependence · copy-aware truth discovery · multi-source data fusion · truthfinder · depen · benchmark datasets · source reliability

Original article: https://reliablecontext.com/articles/truth-discovery-copying-sources-benchmark



In this article

1. Executive Summary
2. Introduction and Background
3. Independence-weighted truth discovery
4. Copy-aware dependence methods
5. Temporal and multi-truth methods
6. Ground-truth-free comparison
7. Feature Comparison
8. Performance and Benchmarks
9. Data Analysis and Evidence
10. Benchmark Design and Method Selection
11. Implications and Future Directions
12. Conclusion

Executive Summary

There is no defensible single winner in a **truth discovery benchmark**. The relevant choice is between assumptions. Independence-weighted methods such as **TruthFinder**, **3-Estimates**, and **Investment** infer source quality from agreement but do not model copying. Dependence-aware methods such as **DEPEN** and **ACCUCOPY** discount evidence from inferred copiers. Temporal models add changing facts and publication delays. Multi-truth models allow several correct values. Ground-truth-free comparators rank methods indirectly, but one prominent proposal, **CompTruthHyp**, assumes sources are independent (arxiv.org). The current survey taxonomy still separates single from multiple values, independent from dependent sources, static from dynamic sources, and supervised from unsupervised learning (researchers.mq.edu.au).

Copying changes the unit of evidence. Five sources repeating one upstream claim are not five independent observations. Dong, Berti-Equille, and Srivastava's dynamic model represented source quality through coverage, exactness, and freshness (vldb.org). In its synthetic tests, copier-identification accuracy exceeded **95%** when coverage exceeded **0.2**, while co-copier accuracy averaged **88%** (vldb.org). These are study-specific results, not transferable operating guarantees.

The benchmark evidence is equally conditional. Waguih and Berti-Equille compared **12 algorithms** and found no clear average precision-versus-runtime benefit over majority voting across five real datasets (arxiv.org). A later axiomatic study describes practical truth-discovery datasets as extremely sparse (link.springer.com). A benchmark should therefore report coverage, conflict, copying topology, truth cardinality, change rate, label sampling, convergence, runtime, and score calibration separately.

The recommended decision rule is conservative: start with majority vote, add model complexity only when observable evidence supports its assumptions, and abstain when dependence or truth cardinality is not identifiable. Preserve every claim, timestamp, method version, parameter, and label decision. W3C notes that provenance can support assessments of quality, reliability, and trustworthiness (w3.org). A production record should keep the selected value and its alternatives auditable rather than treating a source-reliability score as an absolute probability.

Introduction and Background

Truth discovery is the joint task of estimating source reliability and selecting one or more values from conflicting claims. A source can be a website, database, document stream, sensor, annotator, or extraction pipeline. A claim links a source to a value for an object and attribute. **Ground truth** is independently verified reference data used to evaluate the inferred values. TruthFinder's reciprocal premise is representative: trustworthy sources provide many true facts, while facts supplied by trustworthy sources receive more confidence (experts.illinois.edu).

That circular structure is useful but not self-validating. The original TruthFinder article appeared in the June 2008 issue of IEEE Transactions on Knowledge and Data Engineering (experts.illinois.edu). A score may be only a ranking or algorithm-specific weight. Pasternack and Roth warn that generic fact-finder scores typically lack meaningful inherent semantics (aclanthology.org). Treating 0.8 as an 80% probability is unjustified unless the method defines and validates that interpretation. Calibration metrics are appropriate only for outputs with probabilistic semantics. For those outputs, Brier score and log loss are strictly proper scoring rules (link.springer.com).

Eight distinctions control method selection:

- **Independence:** whether matching claims represent separate observations.
- **Dependence:** whether one source copies, republishes, or shares a generator with another.
- **Freshness:** whether a source reports a changed fact promptly, late, or not at all.
- **Truth cardinality:** whether exactly one value or several values may be correct.
- **Supervision:** whether verified labels exist for training, tuning, or evaluation.
- **Coverage:** the proportion of objects or claims observed by each source.
- **Conflict:** the number and distribution of competing values.
- **Abstention:** the option to declare that evidence is insufficient.

These properties are not implementation details. They determine what a vote means. A rigorous comparison must therefore ask which observations make each assumption defensible, not which method has the largest isolated score.

Within **multi-source data fusion**, this report treats truth discovery algorithm comparison as an assumptions audit. It evaluates source-dependence and copy-aware methods, benchmark datasets, evaluation metrics, source copying detection algorithms, and truth discovery method selection under limited evidence.

Absence needs an explicit model as well. The Truth Existence Model represents source behavior with a **silent rate**, **false-spoken rate**, and **true-spoken rate** (pubmed.ncbi.nlm.nih.gov). Its authors evaluated the proposal on three real-world datasets (pubmed.ncbi.nlm.nih.gov). A benchmark that converts silence to a negative claim evaluates a different problem.

Independence-weighted truth discovery

Capabilities

Majority voting is the essential baseline. A reference implementation assigns every source trust 1 and normalizes support by source count (truthdiscovery.readthedocs.io). It has no learned source quality, dependence model, temporal state, or calibrated confidence. Its simplicity is a virtue because a complex method should demonstrate incremental value against it.

TruthFinder iteratively reinforces source trust and fact confidence. **Cosine**, **2-Estimates**, and **3-Estimates** similarly refine truth and trust to a fixed point (pierre.senellart.com). 3-Estimates extends 2-Estimates with an estimate of fact difficulty (pierre.senellart.com). **Investment** distributes each source's trust across its claims before returning credit to sources (aclanthology.org). PooledInvestment adds competition within mutually exclusive claim sets.

Evidence base

The original Galland et al. synthetic experiment reported global precision of **84.5% for Voting**, **88.1% for 2-Estimates**, **88.2% for Cosine**, and **91.5% for 3-Estimates** in its typical setting (pierre.senellart.com). The same paper found 2-Estimates unstable and sometimes worse than baselines (pierre.senellart.com). Both findings belong to that experimental design.

Strengths and limitations

These methods are reasonable when source errors are approximately independent, values are static during the observation window, and a single truth is expected. They are weak choices when agreement can be manufactured by a shared upstream source. Their iterative procedures also require explicit stopping rules. The comparative implementation standardized convergence at cosine-similarity change no greater than **0.001** (arxiv.org). Some models still reached the experiment's **500-iteration** cap without convergence (arxiv.org). A benchmark must count non-convergence rather than silently converting the last iterate into a result.

Formal comparisons can also be narrower than operational use. One axiomatic treatment explicitly assumes that sources act independently (orca.cardiff.ac.uk). The assumption makes analysis tractable, but it must remain visible when conclusions are transferred to copied-source networks.

Copy-aware dependence methods

Capabilities

Copy-aware methods treat agreement as evidence whose weight depends on source lineage. The original dependence model distinguishes independent sources from copiers and treats a shared false value as strong dependence evidence because it is unlikely under independence (static.googleusercontent.com). **DEPEN** discounts votes from likely copiers. Related ACCU variants change assumptions about source accuracy, value similarity, and dependence. These variants are not interchangeable labels.

The key inference problem is circular: source accuracy helps identify false-value coincidences, while copy probabilities alter the evidence used to estimate accuracy. ACCUCOPY iterates until source accuracy stabilizes (static.googleusercontent.com). Its original analysis warns that adding learned source accuracy can remove the convergence guarantee available in a simpler case (static.googleusercontent.com).

Evidence base

In the original book case, the reference set contained **100 randomly selected books** (static.googleusercontent.com). ACCUCOPYSIM classified **508 pairs** as likely dependent (static.googleusercontent.com). These figures demonstrate feasibility, not universal superiority.

Global copy detection addresses structures that pairwise decisions cannot distinguish, including co-copying, transitive copying, and copying from multiple sources (lunadong.com). Its AbeBooks case included **877 bookstores**, **1,263 books**, and **24,364 listings**, but lacked a gold standard for that dataset (lunadong.com).

Strengths and limitations

Copy-aware inference is defensible when the dataset contains duplicated rare errors, lag patterns, or known syndication edges. It is not automatically appropriate for all correlation. Pochampally et al. distinguish copying from broader correlation, such as common extraction behavior or complementary coverage (arxiv.org). Dependence estimates should therefore be evaluated as their own predictions, not treated as invisible preprocessing.

Temporal and multi-truth methods

Capabilities

Static models collapse a stream into one conflict set. That discards update order, publication delay, and the possibility that yesterday's value and today's value are both correct for their times. Dong et al.'s dynamic method combines a **Hidden Markov Model** for whether and when copying occurs with Bayesian truth inference over time (vldb.org). **DynaTD** incrementally updates truth and source reliability, visits incoming data once, and models temporal smoothness (cse.buffalo.edu).

For multi-valued attributes, the **Latent Truth Model (LTM)** represents source sensitivity and specificity separately (vldb.org). This supports multiple correct values for attributes such as authorship or membership, where forcing one winner makes recall structurally impossible.

Evidence base

The 2009 restaurant study crawled its sources weekly from January 22 through March 12 (vldb.org). It included **5,269 restaurants** mentioned by at least two sources and phone-verified deleted listings, finding **280 closed** (vldb.org). Snapshot-only baselines produced average F1 of **0.25 for CURR** and **0.16 for INIT** in its synthetic setup (vldb.org).

DynaTD's reported stock benchmark used **1,000 stocks from 55 sources** on weekdays in July 2011, while its flight benchmark used **1,200 flights from 38 sources** in December 2011 (cse.buffalo.edu).

Strengths and limitations

Temporal models require timestamps that reflect observation or publication, not merely collection. They also require a definition of identity stable enough to compare versions. The 2009 model assumes schema matching and object resolution have already assigned each object a unique identity (vldb.org). Multi-truth models require evaluation at set level, with precision and recall, rather than single-winner accuracy alone.

Ground-truth-free comparison

Capabilities

CompTruthHyp addresses method comparison when reference truth is unavailable. It scores how well a method's inferred outputs can probabilistically explain observed source-value claims (arxiv.org). It covers both single-valued and multi-valued objects (arxiv.org).

Evidence base

The synthetic design fixed **50 sources**, **1,000 objects**, and **20 distinct values per object** (arxiv.org). Its refined Book-Author reference data covered **7.91% of objects** (arxiv.org).

Strengths and limitations

Ground-truth-free selection is useful as a diagnostic, not a substitute for external validation. CompTruthHyp assumes source independence and acknowledges that copying relations were ignored by it and the compared methods (arxiv.org). Its empirical study is evidence for the proposal, not an independent replication. This review located no explicit independent reproduction of its complete validation or of the exact 2009 copying thresholds. That negative search result should be read as a literature-audit limitation, not proof that no reproduction exists.

Feature Comparison

Table 1 compares candidate classes by the assumptions a reviewer must defend. “Score” describes output semantics claimed by the cited method, not probability calibration.

Candidate	Independence and copying	Time and truth cardinality	Supervision	Output and best use
Majority vote	Assumes each vote counts equally; no copying model (truthdiscovery.readthedocs.io)	Static, normally one winner	None	Normalized support; mandatory baseline
TruthFinder	Independence-weighted reinforcement	Static, usually single truth	None	Algorithm-specific fact confidence and source trust (experts.illinois.edu)
3-Estimates	No explicit copying model	Static; adds fact hardness	None	Fixed-point scores; useful when difficulty varies (pierre.senellart.com)

Candidate	Independence and copying	Time and truth cardinality	Supervision	Output and best use
Investment family	No explicit copying model	Static, mutually exclusive claims in pooled form	None	Trust allocation and returned credit (aclanthology.org)
DEPEN/ACCUCOPY	Infers and discounts copying	Static snapshots	None	Value confidence, source accuracy, dependence estimates (static.googleusercontent.com)
Dynamic HMM copying	Infers whether and when copying occurs	Evolving single values and delays	Labels used for evaluation, not required for inference	Time-indexed copying and truth estimates (vldb.org)
LTM	Typically independent-source model	Static multi-truth	Optional priors	Sensitivity, specificity, and fact probabilities (vldb.org)
CompTruthHyp	Explicit independence assumption	Static, single and multi-valued	No ground truth required	Comparative method ranking, not object-level truth production

For algorithm selection, the decisive column is the assumption that matches the observed source network.

Performance and Benchmarks

The strongest broad comparison remains informative precisely because it did not identify a universal winner. Waguih and Berti-Equille evaluated synthetic data and **five real-world datasets**, measuring runtime, iterations, memory, and quality (arxiv.org). Their generator varied coverage, conflict, and positive-claim distributions. Performance differed by dataset. That pattern supports scenario selection, not an overall league table.

Runtime can be part of the statistical question. In one scalability configuration, dependence-aware methods required **5,492 seconds for DEPEN** through **5,788 seconds for ACCUSIM**, while faster non-dependence methods were below **16 seconds** (arxiv.org). The tested scale extended to **10,000 items and 10,000 sources**, while a 100-million-claim combination exceeded available memory (arxiv.org). Hardware, implementation, and stopping criteria prevent direct transfer of those timings.

A benchmark should distinguish eight metric families:

- **Object truth:** accuracy for single truth; precision, recall, and F1 for multi-truth sets.
- **Source quality:** ranking correlation, error against labeled reliability, or calibration where semantics permit.
- **Dependence recovery:** edge precision, edge recall, F1, direction accuracy, and topology-specific results.
- **Temporal recovery:** value accuracy by time, change-point delay, freshness error, and stale-value duration.
- **Operational cost:** wall time, peak memory, iterations, non-convergence rate, and incremental update cost.

- **Robustness:** performance across coverage, conflict, copying, label noise, and initialization.
- **Calibration:** Brier score, log loss, and reliability diagrams only for declared probabilities.
- **Abstention:** risk-coverage curves and error among answered objects.

The metrics cannot be collapsed without an explicit utility function. An F1 score for detected copy edges does not validate truth probabilities. Likewise, correct truth labels do not prove that inferred dependence edges reflect real copying.

Simple modern baselines remain valuable. A 2023 study proposed average proximity to other workers as a competence estimator (ojs.aaai.org). A separate calibration study described an accuracy-preserving truth estimator with a no-accuracy-drop guarantee (proceedings.mlr.press). These are useful comparators when a complex source model claims practical benefit. Even for probabilities, conventions matter: one calibration survey's multiclass Brier definition ranges from **0 to 2**, while half that quantity is also commonly called the Brier score (link.springer.com).

Data Analysis and Evidence

Table 2 separates reported measurements from what they establish. Every row is paper-reported; this review found no independent reproduction of the exact headline thresholds.

Study and setup	Reported result	What the result supports	What it does not support
Galland et al. synthetic typical setting	Voting **84.5%**, 2-Estimates **88.1%**, Cosine **88.2%**, 3-Estimates **91.5%** global precision (pierre.senellart.com)	Performance under that generator	A universal ranking
Dynamic copying synthetic Category II	Copier accuracy above **95%** at coverage above **0.2**; co-copier accuracy **88%** on average (vldb.org)	Coverage and topology affect detectability	Accuracy on a new source graph
Dynamic restaurant study	COPYCEF precision **0.86**, recall **0.87**, F-measure **0.86**; **5,186** ever-existing restaurants inferred (vldb.org)	Joint time and copying evaluation in that case	General calibration
LTM Book benchmark	**1,263 books**, **2,420 facts**, and **48,153 claims** (vldb.org)	Multi-truth scale and partial labeling	Full-population accuracy
Waguih and Berti-Equille Flight	Gold standard representing **7.76%** of complete ground truth (arxiv.org)	Evaluation on a sampled reference	Unbiased full-data estimates without sampling analysis
Waguih and Berti-Equille Biography	Gold standard representing **0.069%** of complete ground truth (arxiv.org)	A very sparse label regime	Fine method separation
Truth Existence flight data	**2,909 flights**, **17,310 questions**, **341,732 non-empty answers**, plus **1,596 no-truth questions** and **80,949 empty answers** (cse.buffalo.edu)	Explicit evaluation of absence	That empty always means false

The dominant quantitative pattern is not a winner but **sensitivity to experimental construction**. Ground-truth coverage ranges from fractions of a percent to small manually labeled samples. The 2014 paper notes that sampled real-data gold standards in prior work were generally under **10%** of the source dataset (arxiv.org).

Label selection must therefore be treated as a sampling design. Confidence intervals, object strata, and adjudication disagreements are more informative than a single point score.

An illustrative vote shows why dependence matters. Suppose sources A through E copy the same upstream claim **X**, while independent sources F and G report **Y**. Majority vote gives X five votes and Y two. If the five-source cluster is conservatively treated as one effective observation, X has effective support one and Y has two. The example is not an empirical estimator. It demonstrates that the answer can reverse solely because the benchmark changes its unit from source pages to independent evidence lineages.

Benchmark Design and Method Selection

A minimum benchmark should have both controlled stress tests and a real-data audit. Synthetic data identifies causal failure modes because coverage, conflict, copying, and temporal change can be varied independently. Real data tests whether those factors were modeled realistically. Neither half substitutes for the other.

Table 3 specifies controllable factors, measurements, and failure signatures.

Factor	Controlled levels	Report separately	Failure signature
Coverage	Per-source and per-object sparsity; overlap graph	Truth score by coverage decile	A method wins only on dense objects
Conflict	Number of values; true-value prevalence; rare false values	Accuracy or set F1 by conflict count	Majority dominates as conflict distribution changes
Copying	Direction, depth, fan-out, co-copying, partial independence	Truth metrics and edge-recovery metrics	High truth score with incorrect lineage
Time	Change rate, delay, crawl interval, missing snapshots	Time-indexed accuracy and detection lag	Stale values score as current truth
Truth cardinality	Zero, one, or multiple correct values	Accuracy for one; precision, recall, F1 for sets	Forced single winner suppresses recall
Labels	Fraction, stratification, adjudicator noise, temporal validity	Confidence intervals and disagreement	Rank reversal under a different sample
Optimization	Initialization, seed, tolerance, iteration cap	Runs, convergence rate, variance	Last iterate reported as convergence
Scale	Sources, objects, claims, versions	Time, memory, incremental cost	Accuracy preserved but resource limit exceeded

The factors are also documentation obligations. ACM SIGMOD reproducibility guidance calls for configuration, initialization, scripts, workload, and measurement protocol (reproducibility.sigmodconf.hosting.acm.org). It recognizes that exact runtime can vary by hardware while reproduced behavior should still support the paper's conclusions (reproducibility.sigmodconf.hosting.acm.org).

Generic metadata is not enough. DataCite states that its generic schema is not intended to replace discipline-specific metadata (schema.datacite.org). Truth-discovery artifacts should therefore add copying topology, conflict distribution, observation time, truth cardinality, and label-construction fields.

Minimum protocol

- **Freeze inputs:** retain raw claims, source identifiers, timestamps, and object-resolution version.
- **Define the estimand:** state whether the target is one value, a value set, a time-indexed state, or existence.
- **Map dependence:** document known feeds, syndication, shared extraction, and unresolved correlation.
- **Specify labels:** give the sampling frame, selection probabilities, adjudication protocol, and label date.
- **Separate tuning:** keep hyperparameter selection objects and test objects disjoint where supervision is used.
- **Run a baseline:** include majority vote and a simple source-weighted model.
- **Stress assumptions:** vary copying, coverage, conflict, cardinality, delay, and label noise independently.
- **Track convergence:** record tolerance, iterations, restarts, seeds, failures, and last-iterate handling.
- **Measure resources:** report hardware, wall time, memory, batch size, and update mode.
- **Preserve provenance:** connect every output to input version, method version, parameters, and label set.
- **Report uncertainty:** provide intervals or repeated-run dispersion with their construction method.
- **Test abstention:** allow no answer when evidence or model identification is inadequate.

W3C's PROV Data Model defines derivation as transformation from one entity into another, a useful relation for tracing benchmark outputs to inputs (w3.org). DataCite describes metadata provenance as recording what changed, when, and by whom (support.datacite.org). Provenance does not make a value true, but it makes a benchmark result reproducible and contestable.

Decision tree

- **If copied sources are implausible and facts are static:** begin with majority vote and independence-weighted methods.
- **If rare shared errors or known feeds exist:** compare DEPEN-family methods with the independent baseline.
- **If update order is observed:** use a temporal dependence model and score change recovery separately.
- **If several values may be correct:** use a multi-truth model and set-valued metrics.
- **If no value may exist:** model silence or non-existence instead of treating every blank as false.
- **If ground truth is sparse:** stratify labels and report uncertainty; do not extrapolate silently.
- **If ground truth is absent:** use CompTruthHyp or agreement diagnostics only as provisional selection evidence.
- **If scores are uncalibrated:** report rankings or relative confidence, not probabilities.
- **If identity is unstable:** resolve and version object identity before comparing truth methods.
- **If assumptions cannot be distinguished:** abstain, preserve alternatives, and request targeted verification.

NIST requires the basis for a probability interpretation or confidence level attached to an uncertainty statement to be given (nist.gov). Recent NIST benchmark work also warns that common metric analyses may produce invalid uncertainty estimates or hide assumptions (nist.gov). Those principles apply directly to truth-discovery

scores and rank comparisons.

Implications and Future Directions

The hardest open problem is **identifiability**. Agreement can arise from independent observation, direct copying, a common upstream feed, shared extraction software, or a genuinely easy fact. Claim matrices alone may not distinguish them. A recent dynamic-dependence study names partial independence as a distinct state, underscoring that sources may mix copied and self-reported claims (web.science.mq.edu.au). Benchmarks need source-network evidence beyond value agreement.

The second challenge is **transfer**. Error rates and copying patterns learned from one attribute or period may not transfer to another. Generalized Fact-Finding argues that conventional methods omit contextual evidence such as extraction confidence and source attributes (cogcomp.seas.upenn.edu). Context can improve inference, but it also increases the burden of documenting features and preventing leakage.

The third challenge is **limited labels**. Extremely sparse reference data can make small score differences non-identifiable. An axiomatic analysis describes practical truth-discovery datasets as extremely sparse and argues that results can apply narrowly to particular algorithms (link.springer.com). Future work should prioritize active selection of verification targets, calibration under selective labels, and evaluation across changing source graphs.

Sparse overlap also creates a minimum-data problem. An empirical-Bayes analysis notes that competence estimation in the real-valued algorithms it studied requires more than one source (proceedings.mlr.press). Benchmarks should report disconnected or singly covered objects instead of allowing evaluation to hide them.

Conclusion

A truth discovery benchmark should select methods by **observable source-network assumptions**, not by a universal ranking. Majority vote is the baseline. Independence-weighted methods add learned trust when copying is negligible. DEPEN-family methods are justified when dependence signals exist. Temporal models are necessary when values and copying relations change. Multi-truth models are necessary when several answers can be correct. Ground-truth-free comparison is provisional and must disclose its own assumptions.

The published evidence shows substantial conditional gains, but it also shows sparse reference labels, convergence failures, large runtime differences, and scenario-specific rankings. High paper-reported copying accuracy is evidence about one synthetic regime, not a portable guarantee. The absence of a reproduced universal winner is not a defect in the field. It is the reason benchmark design must expose coverage, conflict, topology, time, cardinality, label construction, uncertainty, and computational cost.

The defensible workflow is therefore simple in principle: preserve claims and provenance, start with a transparent baseline, add only the complexity supported by observed structure, measure each inferred component separately, and abstain when evidence cannot identify the model. A system of record can preserve the selected value, alternatives, source text, dates, and prior states. It cannot convert an untested assumption into truth. That distinction is the core requirement for researchers and context engineers who must defend every value on which an organization acts.

Source links

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

1. https://arxiv.org/pdf/1708.02029
2. https://researchers.mq.edu.au/en/publications/a-survey-on-truth-discovery-concepts-methods-applications-and-opp
3. https://www.vldb.org/pvldb/vol2/vldb09-335.pdf
4. https://arxiv.org/pdf/1409.6428
5. https://link.springer.com/article/10.1007/s10458-022-09569-3
6. https://www.w3.org/TR/prov-overview
7. https://experts.illinois.edu/en/publications/truth-discovery-with-multiple-conflicting-information-providers-o
8. https://aclanthology.org/C10-1099.pdf
9. https://link.springer.com/article/10.1007/s10994-023-06336-7
10. https://pubmed.ncbi.nlm.nih.gov/26705507
11. https://truthdiscovery.readthedocs.io/en/latest/algorithms.html
12. https://pierre.senellart.com/publications/galland2010corroborating.pdf
13. https://orca.cardiff.ac.uk/id/eprint/150628/8/Singleton-Booth2022_Article_TowardsAnAxiomaticApproachToTr.pdf
14. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41657.pdf
15. https://lunadong.com/publication/global_vldb.pdf
16. https://arxiv.org/abs/1503.00306
17. https://cse.buffalo.edu/%7Ejing/doc/kdd15_evolve.pdf
18. https://vldb.org/pvldb/vol5/p550_bozhao_vldb2012.pdf
19. https://ojs.aaai.org/index.php/AAAI/article/view/25750
20. https://proceedings.mlr.press/v161/ma21a.html
21. https://cse.buffalo.edu/%7Ejing/doc/kdd15_Existence.pdf
22. https://reproducibility.sigmodconf.hosting.acm.org
23. https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf
24. https://www.w3.org/TR/prov-dm
25. https://support.datacite.org/docs/tracking-provenance
26. https://www.nist.gov/pml/nist-technical-note-1297/nist-tn-1297-7-reporting-uncertainty
27. https://www.nist.gov/publications/expanding-ai-evaluation-toolbox-statistical-models?trk=article-ssr-frontend-pulse_little-text-block
28. https://web.science.mq.edu.au/%7Eqsheng/papers/TKDE-2026-truth-discovery.pdf
29. https://cogcomp.seas.upenn.edu/papers/PasternackRo11.pdf
30. https://proceedings.mlr.press/v180/ben-shabat22a.html

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.