Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Google Knowledge Vault: What Survived at Web Scale

> 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 · 2026-09-23 · probabilistic knowledge bases · knowledge graphs · knowledge provenance · knowledge base construction · web-scale knowledge extraction · knowledge graph quality

Original article: https://reliablecontext.com/articles/knowledge-vault-lessons-web-scale



In this article

1. Executive Summary
2. Introduction and Background
3. Definition and Terminology Boundaries
4. Key Changes in the Knowledge Vault Lineage
5. Data Analysis and Evidence
6. Implementation Considerations and Process Changes
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

Google Knowledge Vault was a 2014 research system for constructing a probabilistic knowledge base from web extractions and prior knowledge. It was not an extract, transform, load product, a model-merging technique, or a synonym for Google Knowledge Graph. The KDD paper described supervised fusion across text, tables, document object model structure, human annotations, and prior repositories (research.google). Its companion empirical study, *From Data Fusion to Knowledge Fusion*, analyzed a snapshot dated **October 2, 2013** containing **6.4 billion extracted triples** that collapsed to **1.6 billion unique triples** (gabrilovich.com). Those figures are historical paper results, not a current Google product benchmark.

The scale remains instructive, but the evaluation design is the more durable result. The corpus combined **12 extractors**, more than **1 billion web pages**, and roughly **43 million entities** (gabrilovich.com). The authors reported that **83%** of unique extracted triples were absent from Freebase and estimated raw accuracy at about **30%**, under an assumption that unlabeled triples resembled labeled ones (gabrilovich.com).

For a new study, **reuse** the separation among extraction, provenance, prior knowledge, and calibrated fusion. **Re-test** source dependence, provenance granularity, predicate-specific behavior, long-tail reliability, and probability calibration on independent splits. **Reject** the assumptions that absence means false, one global source score is enough, every predicate has one truth, or a web-derived result transfers to private enterprise evidence. Later work supports these cautions: closed-world evaluation can rank stronger knowledge graph completion models below weaker ones (papers.nips.cc), and open-world calibration requires unknown predictions to remain unlabeled until evidence is obtained (aclanthology.org). This article does not make a finding about independent reproduction of the original experiment.

Introduction and Background

The phrase **knowledge fusion** is overloaded. In the Knowledge Vault lineage, it means estimating whether candidate subject, predicate, object triples are true by considering the provenance of web extractions and, in the KDD system, prior graph knowledge. The PVLDB paper defines the output directly as a probability for each unique triple (www.vldb.org). This is narrower than the broad 2020 survey usage, which also encompasses ontology matching, record linkage, entity resolution, and entity alignment (link.springer.com).

Two publications establish the 2014 record. *Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion* appeared at KDD 2014 on pages 601 to 610 (research.google). *From Data Fusion to Knowledge Fusion* was a separate PVLDB article by seven overlapping authors (gabrilovich.com). The former presents the broader architecture and supervised use of graph priors. The latter adapts Vote, Accu, and PopAccu to the extracted corpus and gives the detailed error analysis. Calling either one a reproduction of the other would be inaccurate.

The central decision in 2026 is therefore not whether to copy a named architecture. It is which empirical propositions still deserve a place in a new benchmark. A defensible study needs to separate what the papers measured from what later surveys inherited, and it must preserve unknown states rather than manufacture negatives from missing graph edges.

That separation changes the research question. Scale alone asks whether a system can process a large candidate set. Fusion quality asks whether its scores survive changes in source mix, predicate semantics, label coverage, and time. Operational defensibility asks a third question: whether a reviewer can reconstruct why one value was retained, what alternatives existed, and what was unknown. This report evaluates the lineage against all three questions.

Definition and Terminology Boundaries

Knowledge fusion begins after candidate claims exist, but many neighboring tasks can occur before or after it. **Extract, transform, load (ETL)** combines, cleans, organizes, and loads data from multiple sources (ibm.com). **Entity alignment** finds descriptions in different graphs that denote the same real-world entity (link.springer.com). **Knowledge graph completion (KGC)** predicts missing links in an incomplete graph (www.ijcai.org). **Model fusion** combines parameters or predictions of models, not factual claims and their sources (arxiv.org).

Table 1 separates the search intents that are most often conflated.

Term	Primary object	Core decision	Not sufficient for
ETL and data integration	Records, columns, and files	How should heterogeneous data be transformed and loaded?	Deciding which contradictory claim is true
Entity alignment or record linkage	Entity descriptions	Do two descriptions refer to the same real thing?	Selecting among conflicting values after identity is resolved
Knowledge fusion	Claims with provenance	Which claims should be believed, with what uncertainty?	Discovering every missing link in a graph
Truth discovery	Conflicting values and sources	What values are true, and how reliable are the sources?	Extraction and semantic linkage by itself
Knowledge graph completion	Existing graph topology and features	Which absent edges are plausible?	Establishing that a web source actually asserted an edge
Model fusion or merging	Model weights or predictions	How should several learned models become one?	Maintaining source-level factual provenance

The boundary matters experimentally. Truth discovery can be treated as the conflict-resolution stage of integration, producing merged records from contradictory sources (vldb.org). Entity alignment is upstream identity work, while link prediction within a single graph is explicitly a different task (link.springer.com). A benchmark that mixes these labels cannot tell whether an error came from extraction, linkage, conflict resolution, or completion.

Key Changes in the Knowledge Vault Lineage

From a graph of assertions to a graph of probabilities

The KDD architecture contained **three major components**: extractors, graph-based prior models, and knowledge fusion (www.cs.ubc.ca). Candidate facts came from free text, HTML document object model trees, and web tables (www.cs.ubc.ca); Google's primary description also identifies human annotations (research.google). The final estimate conditioned on extractions and prior edges (www.cs.ubc.ca). This decomposition remains useful because it prevents a high-confidence extraction from being confused with an independently corroborated fact.

The KDD implementation was supervised. It fit a separate classifier for each predicate, used boosted decision stumps after they outperformed an initial logistic model, and applied Platt scaling to calibrate scores (www.cs.ubc.ca). Reusing that conceptual separation is defensible. Assuming that the same classifier family, predicate partitioning, or calibration mapping remains optimal is not.

From source agreement to provenance-aware reliability

The PVLDB formulation made the input explicitly three-dimensional: data item, web source, and extractor (www.vldb.org). For adaptation to conventional data-fusion algorithms, an extractor and URL pair became a source (www.vldb.org). Refining provenance further to extractor, site, predicate, and pattern reduced weighted calibration deviation by **13%** in that experiment (www.vldb.org).

This finding survives as a design hypothesis, not a universal constant. W3C PROV defines provenance through the entities, activities, and people involved in producing data (www.w3.org). Database-provenance research further distinguishes fine-grained tracking at the level of individual data items (www.ipvs.uni-stuttgart.de). A 2026 study should compare provenance resolutions directly and charge their storage, labeling, and inference costs to the evaluation.

From one truth to multiple, hierarchical, and changing truths

The original fusion assumptions were stressed by the data itself. The PVLDB paper reported that **72% of predicates** were non-functional and **76% of data items** used non-functional predicates (www.vldb.org). Later multi-truth work jointly asks how many truths exist and what they are (arxiv.org). Hierarchical truth discovery adds that even values of nominally functional predicates can sit at different levels of generality rather than being mutually exclusive (arxiv.org).

Time adds another axis. Temporal knowledge bases associate triples with times at which the relation is valid (arxiv.org). W3C PROV-O treats revision as a kind of derivation, providing a standard vocabulary for version lineage (www.w3.org). A value can therefore be correct for one interval, superseded later, and still necessary to

reproduce a past state.

Subsequent related work

The 2020 multi-source fusion survey cites Knowledge Vault but presents a taxonomy and future research agenda, not a rerun of the 2014 corpus (link.springer.com). It also states that no unified theoretical system had yet formed (link.springer.com).

Subsequent studies test pieces of the problem. An independent evaluation implemented **12 truth-discovery algorithms** and studied initialization, convergence, parameters, and scale, but not the original Knowledge Vault corpus (arxiv.org). OpenEA surveyed **23 entity-alignment approaches** and included **12** in its library, addressing identity matching rather than triple-truth calibration (vldb.org). DeepDive evaluated incremental probabilistic construction across **five systems**, again a distinct task and corpus (www.vldb.org). The defensible conclusion is extension without direct reproduction.

Data Analysis and Evidence

The two 2014 papers report related but non-identical slices of the system. The KDD paper says Knowledge Vault held **1.6 billion triples**, including **324 million** at confidence at least 0.7 and **271 million** at confidence at least 0.9 (www.cs.ubc.ca). The PVLDB study reports the extraction corpus and fusion audit in more detail.

Table 2 is an evidence ledger, with denominators and assumptions kept beside each result.

Claim	Reported value and denominator	Measurement or assumption	2026 interpretation
Corpus scale	**6.4B extracted**, **1.6B unique** triples (gabrilovich.com)	Snapshot dated October 2, 2013	Historical workload, not a current benchmark
Extraction breadth	**12 extractors**, over **1B pages**	Four text, five DOM, two table, one annotation extractor (www.vldb.org)	Preserve extractor identity in provenance
Gold-standard coverage	**650M of 1.6B**, or **40%**, labeled	Local closed-world Freebase labels (www.vldb.org)	Missing labels are selection, not random truth
Positive gold labels	**200M of 650M** labeled triples	Freebase match counted as correct (arxiv.org)	Class balance depends on the reference graph
Extractor accuracy	**0.09 to 0.78**	Observed under the local gold standard (arxiv.org)	Report by extractor and predicate
Evidence sparsity	**75%** from one extractor; **51%** from one page	Unique-triple denominators (arxiv.org)	Agreement cannot rescue most of the long tail
POPACCU+ high score	**0.94 accuracy** above a 0.9 score (gabrilovich.com)	Observed against paper gold labels	Calibration is conditional on label validity

The KDD paper says about one third of the **271 million** high-confidence triples were not in Freebase (www.cs.ubc.ca), while PVLDB says **83% of all unique extracted triples** were absent. These use different denominators and are compatible.

Implementation Considerations and Process Changes

A new multi-source study should treat Knowledge Vault as a set of hypotheses. The following controls turn those hypotheses into falsifiable experiments.

Dataset and task definition

- **Freeze the snapshot.** Publish source-observation times and dataset hashes.
- **Separate stages.** Score extraction, entity linkage, predicate linkage, and fusion independently.
- **Define the unit.** State whether the unit is an extraction, unique triple, data item, entity, or source.
- **Preserve provenance.** Retain extractor, document, site, predicate, pattern, and observation time when permitted.
- **Declare source policy.** Record which private, licensed, and public sources were eligible.
- **Represent unknowns.** Use explicit true, false, unknown, blocked, and not-applicable states where the task requires them.
- **Model multiplicity.** Mark predicates as functional, set-valued, hierarchical, or time-varying based on a published protocol.
- **Keep identity separate.** Do not count an entity-resolution correction as a fusion improvement.

Gold-standard construction

- **Sample beyond known entities.** Avoid conditioning the entire test set on coverage in one reference graph.
- **Stratify the audit.** Sample by predicate, source frequency, extractor, score band, and observed conflict.
- **Use adjudication.** Record annotator disagreement and a resolution rule.
- **Allow multiple truths.** Label complete value sets, not only one object per subject-predicate pair.
- **Capture hierarchy.** Distinguish contradiction from compatible levels of generality.
- **Date the label.** State the valid time and the time the label was observed.
- **Audit presumed negatives.** Manually inspect absent-reference and hard-negative cases.
- **Publish denominators.** Report every rate with its sampled and population denominator.

Open-world evaluation is essential. NeurIPS 2022 defines the closed-world error as treating all unknown triples as false (papers.nips.cc). InferWiki instead supplies manually annotated negative and unknown triples (aclanthology.org). Its authors also showed that a conventional corruption procedure could produce almost **100% triple-classification accuracy**, illustrating how an easy negative generator can swamp the intended reasoning task (aclanthology.org).

Models, baselines, and ablations

- **Start with transparent baselines.** Include vote, per-source accuracy, and a calibrated supervised model.
- **Ablate priors.** Compare extractions alone, priors alone, and their combination.
- **Ablate provenance.** Compare URL, site, extractor, predicate, and pattern resolutions.
- **Test dependence.** Add copied-source clusters or a dependence-aware baseline.

- **Test long tails.** Report sources with one, two to five, six to twenty, and more than twenty claims.
- **Vary source granularity.** Check whether site-level scores hide page-level or section-level differences.
- **Separate source error.** Distinguish an incorrect extraction from an accurately extracted false assertion.
- **Use temporal splits.** Train on earlier observations and test on later ones.

Knowledge-Based Trust makes the extraction-versus-source distinction explicit (arxiv.org). Long-tail research likewise notes that most sources provide only a few claims and therefore adds confidence intervals around reliability estimates (experts.illinois.edu).

Metrics and reporting

- **Report ranking.** Use precision-recall curves and per-stratum retrieval measures.
- **Report proper scores.** Include Brier or logarithmic score for probabilities.
- **Report calibration.** Supply reliability plots with sensitivity to binning.
- **Report coverage.** Measure the fraction of candidates that receive a decision at each threshold.
- **Report selective risk.** Show accuracy as abstention increases.
- **Report error taxonomies.** Separate extraction, linkage, dependence, multiplicity, hierarchy, time, and label errors.
- **Report uncertainty.** Include confidence intervals and variation across runs.
- **Report costs.** Count annotation, storage, inference, and source-acquisition costs.

Calibration means that predicted probabilities match observed frequencies (www.pnas.org). The Brier and logarithmic scores are strictly proper, but Brier loss alone mixes calibration with resolution and uncertainty (www.pnas.org) (scikit-learn.org). Calibration data should also be independent of training data (scikit-learn.org).

Table 3 maps evaluation considerations to controls and decision criteria.

Failure mode	Observable symptom	Required control	Decision rule
Extraction error	Wrong entity, predicate, or value span	Stage-specific labeled sample	Reject fusion gains that come from changing extraction
Entity or predicate linkage	Claim attached to the wrong graph node	Frozen alignment set and separate score	Report linkage and fusion independently
Source dependence	Copied claims appear as corroboration	Cluster-aware or dependence-aware baseline	Do not count dependent copies as independent votes
Long tail	Unstable reliability for sparse sources	Claim-count strata and intervals	Suppress unsupported point estimates
Temporal drift	Old truth scored against a current label	Valid-time and observation-time labels	Reproduce scores for each evaluation date

The table makes evaluation a chain of custody. Each metric must identify the candidates, labels, source snapshot, and transformation that produced it. W3C PROV can describe entities, activities, and agents, while nanopublications illustrate claim-scale knowledge objects with metadata (nanopub.net). Neither standard

decides truth, but both can make a truth decision inspectable.

Reproducibility package

- **Document assumptions.** State formal restrictions and their expected failure cases.
- **Version everything.** Record code, datasets, ontologies, dependencies, and model artifacts.
- **Record randomness.** Publish seeds and nondeterministic operations.
- **Count runs.** Give the number of runs behind each aggregate.
- **Describe infrastructure.** Report hardware, software, operating system, and library versions.
- **Export decisions.** Provide scores with provenance and label history, subject to source rights.
- **Retain disagreements.** Preserve competing source values and adjudicator notes.
- **Enable replay.** Reconstruct the dataset and labels as they existed at a declared date.

The AAAI-26 reproducibility checklist requires assumptions to be stated clearly, the number of runs to be reported, and variation or confidence information to accompany summaries (aaai.org). NIST additionally warns that library versions, numerical precision, compiler options, and processor architectures can break numerical reproducibility (www.nist.gov).

Implications and Future Directions

Three parts of the Knowledge Vault lineage remain defensible. First, **provenance is part of the prediction input**, not decorative metadata added after scoring. Second, **calibration and ranking answer different questions** and both must be evaluated.

Four assumptions should be rejected before benchmarking. Absence from a reference graph should not automatically become a negative. One truth should not be imposed on set-valued or hierarchical predicates. Repeated extractions should not be treated as independent evidence without a dependence analysis. A probability calibrated on local closed-world labels should not be described as real-world correctness without an open-world audit.

For private and public evidence, the missing benchmark is an end-to-end, rights-respecting test that retains source text, reader permissions, unknown states, identity decisions, and historical versions. The FAIR principles extend beyond data to the algorithms, tools, and workflows that produced it (www.nature.com). The next credible study should therefore publish not only scores but the evidence and process needed to explain every scored claim.

Frequently Asked Questions (FAQs)

What was Google Knowledge Vault?

It was a 2014 research system and paper lineage for constructing a probabilistic knowledge base from web extractions and prior knowledge. Google Research describes supervised fusion of multiple extraction types, not a currently benchmarked commercial service (research.google).

Is knowledge fusion the same as data integration?

No. Data integration and ETL move and harmonize records. Knowledge fusion, in this lineage, estimates the truth of claims using provenance and prior information. A broader survey uses the term for identity alignment and conflict resolution as well, so every study must define its scope.

Was POPACCU+ independently reproduced at Knowledge Vault scale?

This article does not assess whether the original experiment was independently reproduced.

Why is knowledge graph provenance important?

Provenance identifies how a claim was produced and supports review of source dependence, extraction mistakes, and historical state. W3C defines a model in which activities use and produce entities and agents can influence those processes (www.w3.org).

What is the main lesson for a 2026 study?

Benchmark the labels as seriously as the model. Preserve unknowns, audit presumed errors, allow multiple and time-bounded truths, test dependence, and publish denominators.

Conclusion

Knowledge Vault's durable contribution is not a **2014** scale record (research.google). It is a decomposition of web-scale knowledge construction into extraction, provenance, prior knowledge, and probabilistic fusion.

The evidence supports a selective inheritance. Reuse provenance-aware modeling, stage separation, probability calibration, and manual error analysis. Re-test provenance granularity, predicate-specific models, source dependence, and long-tail reliability. Reject closed-world negatives, universal single-truth assumptions, and unqualified transfer from public web pages to private organizational sources.

This standard is deliberately stricter than reproducing a final score. A credible benchmark should make each stage independently inspectable, retain disagreements, and show how conclusions move when labels or source assumptions change. It should reward correct abstention when evidence is unavailable. Most importantly, it should let another researcher distinguish a model improvement from a change in extraction, identity resolution, coverage, or annotation policy.

Source links

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

1. https://research.google/pubs/knowledge-vault-a-web-scale-approach-to-probabilistic-knowledge-fusion
2. https://gabrilovich.com/publications/papers/Dong2014DFK.pdf
3. https://papers.nips.cc/paper_files/paper/2022/hash/378226e5df7eded3e401de5c9493143c-Abstract-Conference.html
4. https://aclanthology.org/2020.emnlp-main.667
5. https://www.vldb.org/pvldb/vol7/p881-dong.pdf
6. https://link.springer.com/article/10.1007/s11280-020-00811-0

7. https://gabrilovich.com/publications/sort_date.html
8. https://www.ibm.com/think/topics/etl
9. https://link.springer.com/article/10.1007/s10618-023-00941-9
10. https://www.ijcai.org/proceedings/2023/734
11. https://arxiv.org/abs/2309.15698
12. https://vldb.org/pvldb/vol5/p550_bozhao_vldb2012.pdf
13. https://www.cs.ubc.ca/~murphyk/papers/kv-kdd14.pdf
14. https://www.w3.org/TR/prov-overview
15. https://www.ipvs.uni-stuttgart.de/departments/de/resources/papers/provenance_survey_online.pdf
16. https://arxiv.org/abs/1705.04915
17. https://arxiv.org/abs/1904.10217
18. https://arxiv.org/abs/2005.05035
19. https://www.w3.org/TR/prov-o
20. https://arxiv.org/abs/1409.6428
21. https://vldb.org/pvldb/vol13/p2326-sun.pdf
22. https://www.vldb.org/pvldb/vol8/p1310-shin.pdf
23. https://arxiv.org/html/1503.00302v1
24. https://aclanthology.org/2021.acl-long.534
25. https://aclanthology.org/2021.acl-long.534.pdf
26. https://arxiv.org/abs/1502.03519
27. https://experts.illinois.edu/en/publications/a-confidence-aware-approach-for-truth-discovery-on-long-tail-data
28. https://www.pnas.org/doi/10.1073/pnas.2016191118
29. https://scikit-learn.org/stable/modules/calibration.html
30. https://nanopub.net
31. https://aaai.org/conference/aaai/aaai-26/reproducibility-checklist
32. https://www.nist.gov/programs-projects/numerical-reproducibility
33. https://www.nature.com/articles/sdata201618
34. https://www.w3.org/TR/prov-dm

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.