Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Knowledge Graph Edge Citation Verification: A Span Test

> 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 · 2026-11-02 · knowledge graph edge citation verification · knowledge graph provenance · source span · llm knowledge graph extraction · hallucinated edges · citation grounding · relation extraction evaluation · graphrag provenance · rdf provenance · attribution evaluation

Original article: https://reliablecontext.com/articles/knowledge-graph-edge-citation-source-span-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes From Triple Plausibility to Source-Span Acceptance
4. The Adversarial Fixture
5. Implementation Considerations and Process Changes
6. Data Analysis and Evidence
7. Frequently Asked Questions (FAQs)
8. Implications and Future Directions
9. Conclusion

Executive Summary

A knowledge graph edge is a compressed claim, and a source citation attached to it is only useful if a reviewer can open the cited text and confirm that it entails the same qualified claim. This report proposes a **source-span acceptance test** for that decision and grounds it in the published literature. The test asks whether the proposed subject, predicate and object, after extraction, referent resolution and canonicalization, are each traceable to a located passage that supports the same claim, including its negation, modality, attribution and time scope.

The need is documented. In an early large-scale pipeline, the authors extracted 97M triples and said that the majority of inconsistencies with a reference graph came from extractor errors, though they reported no error rate (aclanthology.org). A 2026 anchor-based framework traces every triplet element to a character span and measured residual hallucination rates between 0.23% and 20.23% across four backbone models and three datasets (www.mdpi.com). In that study, more than 95% of detected hallucinations were relations unsupported by the text between correctly grounded entities, so entity presence in the passage is not enough (www.mdpi.com).

Provenance vocabularies and review protocols answer different questions. The W3C provenance model records derivation, quotation and attribution, and says provenance supports assessments of reliability rather than establishing them (www.w3.org). The ProVe pipeline tests whether a cited web page supports a Wikidata triple, and it reports 79.4% accuracy over its 416-pair sample, with the higher 87.5% figure limited to text-rich sources (semantic-web-journal.net) (www.semantic-web-journal.net). Attribution research asks annotators to mark output as attributable only if all of it follows from the source (arxiv.org), which is the standard the proposed test adopts.

The report defines six adversarial fixtures (explicit support, negation, hypothetical or future statements, quotation by another speaker, temporal change and coreference failure), plus one canonicalization drift case. It prescribes four verdicts: supported, contradicted, insufficient and inaccessible. It also specifies the record to keep for every verdict and the arithmetic for a support rate and an unsupported-claim rate, with denominators and a Wilson interval (www.afit.edu). A worked hypothetical sample of 200 edges yields 82.0% supported (95% interval 76.1% to 86.7%), illustrating the method, not any measured system.

Evidence limits apply. No published study validates this exact test, none of the cited work measures reviewer time, and the quoted rates use different definitions and datasets, so they are not comparable with each other. The RDF 1.2 representation used in the examples was a Candidate Recommendation snapshot dated 7 April 2026, and it still had that status when rechecked on 2 November 2026; verify the current status before relying on it (www.w3.org).

Introduction and Background

Teams that build knowledge graphs from documents make a recurring decision: may a relation proposed by an extractor enter a graph that supports decisions? The usual guidance is to validate a sample. One widely read construction guide advises starting with a small sample to validate graph structure and data mapping, and tests the finished graph by running queries that answer business questions (neo4j.com). That advice checks the graph model and its usefulness. It does not, on the page read for this report, describe checking an individual extracted edge against the passage it came from.

This report proposes a general evaluation method for that gap. It defines five terms.

- **Assertion candidate:** a (subject, predicate, object) triple, with any qualifiers, proposed by an extractor from one document. An RDF triple has exactly those three components (www.w3.org).
- **Document version:** the exact bytes or snapshot of the source that the extractor read, identified by retrieval date and content hash, because sources change.
- **Source span:** a locator plus verbatim text that identifies the passage offered as support. A text quote selector copies the range of text and optionally records the surrounding prefix and suffix (www.w3.org).
- **Attributed speaker:** the agent who made the statement in the passage, which may differ from the author of the document.
- **Accepted graph claim:** the qualified claim the graph will state, after review, which may be narrower than the candidate (for example "X said Y" instead of "Y").

Scope and evidence limits

The method is a proposal assembled from published components. It is not a benchmark result. Where a number appears, it is attributed to the study that produced it, with that study's scope.

Why the span matters: three published observations

First, provenance is routinely lost. The authors of an early text-to-graph pipeline framed the gap as a graph that cannot answer "how do we know this fact to be true" (aclanthology.org). Second, extraction is not copying. A GraphRAG paper observes that LLM extraction is abstractive, so extracted items may not be explicitly stated in the text (arxiv.org). Third, reviewers need the context. The same early pipeline's authors say that alignment against a reference graph has nuanced cases requiring human judgment, because both the graph and the document can be wrong (aclanthology.org).

Key Changes From Triple Plausibility to Source-Span Acceptance

The proposal changes four habits. Each is described with its supporting literature and its cost.

Change 1: treat the edge as a chain of transformations

An edge is the product of at least three transformations: extraction of a candidate from text, resolution of mentions to canonical entities, and canonicalization of the relation phrase to a schema predicate. Each can change meaning, so each needs its own evidence.

The Extract-Define-Canonicalize framework separates these phases explicitly, as a three-phase process of open extraction, schema definition and post-hoc canonicalization (aclanthology.org). Its motivation is that earlier LLM methods needed the schema in the prompt, which does not scale to large schemas (aclanthology.org). In its target-alignment mode, a language model assesses the feasibility of each candidate transformation, to avoid over-generalization (aclanthology.org). The paper's own baseline comparison shows why this matters: a clustering baseline merged relations such as place of birth and cause of death into one category (aclanthology.org). A graph that stores only the canonical predicate hides that merge.

The earlier pipeline by Clancy, Ilyas and Lin, presented at the Second Workshop on Fact Extraction and VERification in 2019 (aclanthology.org), shows the resolution step. It stored each mention with its class, text and character offsets, and it recorded unlinked mentions explicitly instead of dropping them (aclanthology.org). It also mapped its fixed set of 45 relations to Wikidata properties by hand (aclanthology.org). The anthology page lists the title as "Scalable Knowledge Graph Construction from Text Collections", although the PDF header carries a different title. Each of these is a transformation with its own failure mode.

The acceptance test therefore requires three evidence items per edge: the span supporting the extracted claim, the span or record supporting the referent mapping for subject and object, and the justification that the canonical predicate preserves the meaning of the source phrase. W3C provenance terms can describe the chain: a derivation is a transformation of one entity into another, and a quotation is a particular case of derivation (www.w3.org).

Change 2: attach a locator and verbatim text to every edge

A document identifier is not a span. The established tooling stores provenance at different granularities, and the choice sets the minimum review unit.

- **Entity to chunk links:** the Neo4j GraphRAG Python builder documents relationships between extracted entities and the chunk they were extracted from (neo4j.com). The page does not describe a source locator on each extracted relationship.
- **Document links:** a LangChain integration option stores the source document and links it to nodes (reference.langchain.com).
- **Text units:** Microsoft GraphRAG outputs list the text units a relationship was found within, and a text unit stores the raw full text of the chunk (microsoft.github.io). Its documentation states that extracted items reference text units, which provides provenance back to the original source text (microsoft.github.io). It also notes that larger chunks give less meaningful reference texts (microsoft.github.io).
- **Retrieval with source text:** LlamaIndex's property graph guide shows an option to include the source chunk with matching graph paths at query time (developers.llamaindex.ai).

- **Character-level citations:** a documented LLM API feature returns citations that reference specific locations in supplied documents, with sentence-level chunking that defines the minimum granularity of possible citations (docs.claude.com).

None of these documentation pages reports a measured citation accuracy for edges. A chunk reference tells the reviewer where to look, not whether the passage entails the claim. A character-level locator narrows the search; the entailment judgment remains a separate step.

For web and document sources, the W3C Web Annotation model offers a robust way to write the locator. A position selector records start and end offsets, but the specification warns it is very brittle with regard to changes to the resource (www.w3.org). A quote selector adds a prefix and suffix to distinguish multiple copies of the same sequence of characters (www.w3.org). The model allows several selectors for one segment (www.w3.org), so a record can hold an offset for speed and a quote for resilience. Wikidata's reference practice points the same way: it records the retrieval date and extra properties so the source can be tracked down if the URL changes (www.wikidata.org), and it adds page or section locators because they help when verifying the value (www.wikidata.org).

Change 3: judge entailment of the qualified claim, not entity presence

The central change is the test applied to the span. Lexical presence is a screening step, not acceptance.

The Text2KGBench benchmark defines subject and object hallucination as presence checks: after stemming, a subject or object passes if it appears in the sentence or the ontology concepts, and relation hallucination is an exact match against the ontology's relation list (arxiv.org). These are screening metrics, and they do not test whether the sentence supports the relation.

String matching can also over-reject. A 2026 study of triplet extraction from corporate annual reports reported a strict-regex subject-hallucination rate of 65.2% on one report with one model (arxiv.org). A manual audit found that the flagged cases were instances where the model had correctly captured the entity but changed its surface form (arxiv.org). Both failure directions appear: matching rejects valid edges and accepts coincidental ones. The anchor-based framework concedes that fully restored triplets with coincidental string matches could still be false positives (www.mdpi.com).

Entailment judgments come from the attribution literature. The Attributable to Identified Sources definition asks whether a generic hearer would affirm "According to" the provided parts that the statement holds (arxiv.org). Annotators mark output as attributable only if all of it can be inferred from the source, and misleading wording or changed details count against attribution. The citation-generation benchmark ALCE operationalizes the same idea with a natural language inference model (arxiv.org) and scores recall as 1 only if the concatenation of all cited passages fully supports the segment (arxiv.org). A graph edge with two cited spans should be judged on the concatenation, and each span should be checked for relevance.

The fact-verification tradition supplies the label set. FEVER labels claims as supported, refuted or not enough information; the last applies when the claim cannot be supported or refuted by the evidence available (arxiv.org). SciFact adds rationales, defined as minimal collections of sentences that justify a label (arxiv.org). The concept of "supported by a chosen source, not true" is explicit in FActScore, which says truth is judged

relative to a particular knowledge source (arxiv.org). ProVe makes the same design choice: triples are evaluated not as factual or non-factual, but by whether the reference supports them (semantic-web-journal.net).

Change 4: separate four outcomes from three confounds

The proposed verdicts are:

- **Supported:** the span entails the accepted graph claim, including its qualifiers.
- **Contradicted:** the span, or another span in the same document version, entails the opposite or a conflicting claim.
- **Insufficient:** the span neither supports nor contradicts the claim. This is the analogue of the "not enough information" label.
- **Inaccessible:** the reviewer cannot open the document version or read the text. A scanned PDF without extractable text is one documented example of non-citable content for one citation tool (docs.claude.com).

The verdict is about source support only. Three other questions are recorded separately: whether the extractor was correct about what the passage says (extractor correctness), whether the source is reliable (source reliability), and whether the claim is true in the world. Provenance standards are explicit that they do not settle the last: the W3C provenance model states that provenance can help users make trust judgements (www.w3.org), and the ProVe authors note that Wikidata does not care for truthfulness of information, but rather supportive provenance (semantic-web-journal.net). The Wikibase data model even allows a deprecated statement that faithfully records an erroneous claim made by a cited document (www.mediawiki.org). A supported edge from an unreliable source is still supported; it is routed to source-reliability review.

The Adversarial Fixture

The fixture is a small, labeled set of passages designed so that a naive extractor produces a plausible triple that the passage does not support. It is built from phenomena that the information extraction literature has measured separately: negation and hedging, factuality, attribution, temporal scope and coreference. No published study combines them into one edge-level acceptance test, so the fixture is a design, not a validated instrument.

Evidence for each phenomenon

Negation and speculation. The BioScope corpus was built because detecting uncertain and negative assertions is essential in most biomedical text mining tasks, and it contains more than 20,000 annotated sentences (pmc.ncbi.nlm.nih.gov). The CoNLL-2010 shared task states that an uncertain relation must not be confused with factual textual evidence, and that a single sentence can contain both factual and uncertain parts (aclanthology.org). In its biological evaluation set, 790 of 5003 sentences were uncertain (aclanthology.org).

Modality and factuality. FactBank represents the factual value of an event as a tuple of modality and polarity (www.cs.brandeis.edu). It treats arguments of volition and commitment predicates as possible at a future point in time (www.cs.brandeis.edu). A planned acquisition is not an acquisition.

Attribution. The Penn Discourse Treebank authors report that in preliminary annotation 34% of relations had a non-writer agent as source, and they conclude that attributing everything to the writer would be wrong two-thirds of the time (aclanthology.org). The Quotebank extractor attributes 86.9% of quotations correctly in its experiments, so roughly one in eight speaker attributions in that setting was wrong (dlab.epfl.ch). In W3C terms, attribution is the ascribing of an entity to an agent (www.w3.org), which is the property to record when a passage reports someone else's statement.

Temporality. Wikidata's guidance is that qualifiers can constrain the validity of a value (www.wikidata.org), that correct historical information should be annotated with start and end qualifiers instead of being deprecated (www.wikidata.org), and that qualifiers indicate the period during which a statement is valid (www.wikidata.org). The TimeML authors state that because properties change over time, a database of assertions about entities will be incomplete or incorrect if it does not capture temporal updates (timeml.github.io).

Coreference. The DocRED dataset analysis found that 61.1% of sampled relation instances required reasoning beyond pattern recognition, and the authors state that for instances needing coreference, resolution must be performed first (arxiv.org). Entity linking also loses mentions early: one study found that at least 10% of gold mentions were missed by its mention extractor when mentions were not supplied (aclanthology.org), and it describes metonymy, where a name refers to an associated entity, as a distinct error source (aclanthology.org). The anchor-based framework discussed above works at sentence or passage level without cross-sentence coreference resolution, which the authors list as a limitation (www.mdpi.com).

Fixture design and expected verdicts

Table 1 below sets out the seven fixture cases. All passages are invented for illustration (Hypothetical Example). The entity names are fictional.

Case	Hypothetical source text	Naive candidate triple	Expected verdict and reason	Required action
1. Explicit support	"On 4 March 2025, Halden Labs acquired Brightfield Systems."	(Halden Labs, acquired, Brightfield Systems)	**Supported.** The span states the claim with a date, and both entities are named in the span.	Accept; store span, date qualifier and document version.
2. Negated relation	"Halden Labs did not acquire Brightfield Systems; the talks ended."	(Halden Labs, acquired, Brightfield Systems)	**Contradicted.** Negation reverses the relation, and uncertainty and negation must be separated from factual assertions (pmc.ncbi.nlm.nih.gov).	Reject the edge; keep the span as counterevidence.
3. Hypothetical or future	"Halden Labs said it may acquire Brightfield Systems next year."	(Halden Labs, acquired, Brightfield Systems)	**Insufficient.** The event is possible and future, not factual (www.cs.brandeis.edu).	Abstain, or accept only a narrower claim with a modality qualifier.

Case	Hypothetical source text	Naive candidate triple	Expected verdict and reason	Required action
4. Quotation by another speaker	"A rival's chief executive claimed that Halden Labs acquired Brightfield Systems."	(Halden Labs, acquired, Brightfield Systems)	**Insufficient** for the event; **supported** for "the chief executive claimed the event". The source is attributed to a non-writer agent (www.w3.org).	Correct to a statement-level claim with the speaker as attributed agent.
5. Temporal change	"Until 2023, Mara Quill chaired the board of Brightfield Systems; the role passed to another director in 2024."	(Mara Quill, chairs, Brightfield Systems)	**Contradicted** for the present tense; **supported** for the interval. Facts hold within a period (www.wikidata.org).	Correct with start and end qualifiers and observation date.
6. Coreference failure	"Halden Labs and Brightfield Systems signed in May. The company later said it had ended the agreement."	(Brightfield Systems, ended, agreement)	**Insufficient.** "The company" is ambiguous between two entities; resolution must come first.	Hold; request a span that names the referent.
7. Canonicalization drift	"Brightfield Systems announced a licence agreement with Halden Labs."	(Halden Labs, acquired, Brightfield Systems) after mapping "licence agreement" to an acquisition predicate	**Contradicted or insufficient.** The canonical predicate is broader or different than the source phrase; over-generalization is a documented risk (aclanthology.org).	Reject the mapping; record the source phrase next to the canonical predicate.

The matrix is a test design: each row isolates one transformation or linguistic phenomenon, so an acceptance procedure that passes row 1 and fails rows 2 to 7 is behaving like a presence check. A procedure that fails row 1 is over-rejecting, a failure the string-matching study above shows is common. In practice each case should be instantiated many times with different entities, tenses and sentence structures, and the labeled fixture should be frozen so that extractor or prompt changes can be compared on the same inputs.

Implementation Considerations and Process Changes

The record to keep for each verdict

Each reviewed edge needs a record that a second reviewer can re-open. The Wikidata model is a useful baseline: a statement is a claim plus a list of references, and that list is allowed to be empty (www.mediawiki.org). Wikidata also says statements imported from a Wikimedia project are not considered sourced (www.wikidata.org). The proposed record is stricter. It does not allow an empty evidence field, and it separates the following items.

- **Assertion:** subject, predicate, object and qualifiers as proposed, plus the accepted claim after correction.
- **Passage locator:** an offset pair and a quote selector with prefix and suffix, tied to a document version.
- **Verbatim text:** the span as read, copied exactly.
- **Referent mapping:** the surface mention, the resolved entity and the evidence for the resolution.

- **Predicate mapping:** the source phrase and the canonical predicate chosen.
- **Modifiers:** negation, modality, attributed speaker, temporal scope.
- **Counterevidence:** any span in the same version that conflicts.
- **Observation date and document version:** when the text was read, and a content hash or archive reference.
- **Reviewer, verdict, reason and action:** accept, correct, abstain or reject, with a reason.

In RDF 1.2 terms, a triple term can be used in a reifying triple without asserting the triple: an unasserted triple term is not claimed to be true unless the same triple is also asserted (www.w3.org). The specification also notes that several reifiers can relate to the same proposition, such as statements with different sources (www.w3.org). That makes a candidate recordable before acceptance, with evidence attached per source. The RDF 1.1 Recommendation dates from 25 February 2014 (www.w3.org), while RDF 1.2 Concepts was a Candidate Recommendation snapshot when read, with the page stating that it was not expected to advance to Recommendation any earlier than 5 May 2026, a date that has since passed; the snapshot dated 7 April 2026 still showed Candidate Recommendation status when rechecked on 2 November 2026 (www.w3.org). Verify the current status of RDF 1.2 before depending on it. The W3C provenance ontology is a Recommendation of 30 April 2013 (www.w3.org) and records derivation and primary source relations, noting that secondary sources should reference the primary sources they derive from (www.w3.org). It is a vocabulary for the chain, not a verifier of the passage.

Reviewer procedure

The attribution protocol offers a useful two-stage design. Annotators first judge whether the output is interpretable without the source, and only then see the source to judge support. The source is withheld in the first stage so that evaluators remain strict about ambiguity such as unresolved pronouns (arxiv.org). Applied to edges, the first stage checks the candidate and the referent mapping; the second checks entailment against the span.

1. **Freeze the version.** Fetch and hash the document, and record the observation date.
2. **Check interpretability.** Does the candidate name its entities unambiguously, without reading the source?
3. **Open the span.** Compare the verbatim text with the exact graph claim, qualifiers included.
4. **Search for counterevidence.** Scan the same version for negation, retraction or later change.
5. **Assign a verdict and action.** Accept, correct (narrow the claim, add a qualifier, add an attributed speaker), abstain, or reject.
6. **Double-review a sample.** Measure agreement before trusting the verdict distribution.

A reference label is not ground truth: in one benchmark, human-written reference summaries were attributable to their documents in only 54% of interpretable examples (arxiv.org).

Granularity and cost

Finer spans make review cheaper but extraction harder. A documented LLM citation feature notes that asking for JSON-only output can yield few or no citations, and its recommended parsing step checks that each field has a value and at least one citation (docs.claude.com). A cited offset produced by a model should itself be

checked: the quote at the offsets should equal the quoted text in the document version, which is a cheap deterministic test before any human reads the span.

The anchor-based framework shows the pattern at scale: it requires character-level positions because positions are what provenance tracking depends on (www.mdpi.com), routes fully restored triplets directly to acceptance and partially restored ones to targeted LLM verification (www.mdpi.com). A retrieval and entailment gate in another study raised precision on an open IE benchmark from 64.5% to 75.5% but lowered recall from 74.2% to 57.2% (ceur-ws.org). Verification trades coverage for precision; the abstention outcome exists so that this trade is explicit.

Corrections, abstentions and drift

Three actions need rules in advance. A **correction** narrows the claim so that the span entails it (for example from "acquired" to "said it may acquire"). An **abstention** keeps the candidate out of the decision graph while retaining it with its reason, so a later source can resolve it. A **re-check** is triggered when the document version changes or the observation date passes a policy age (a policy choice proposed in this report). Because properties change over time (timeml.github.io), an accepted edge is a statement about a version at a date, not a permanent fact.

Data Analysis and Evidence

This section collects the measured numbers that bear on the design, each with its scope. They come from different tasks, so they are context for the method, not a ranking of systems.

Table 2 summarizes published measurements on label quality, evidence retrieval and extraction support.

Study (year)	What was measured	Reported result and scope
Alt et al. (2020), TACRED	Re-annotation of the most challenging dev and test examples	Label errors account for 8% absolute F1 test error, and more than 50% of examined examples needed relabelling (arxiv.org).
Stoica et al. (2021), Re-TACRED	Full re-annotation of TACRED	New labels disagree with the original in 23.9% of sentences (arxiv.org).
Tan et al. (2022), Re-DocRED	Missing relations in DocRED	About 64.6% of triples were missing from the original dataset (arxiv.org).
Amaral et al. (2023), ProVe	Support of Wikidata triples by their cited web pages	79.4% accuracy and 57.4% F1-macro over the whole 416-pair sample (semantic-web-journal.net). The higher 87.5% accuracy is for text-rich sources with explicit natural-language support (www.semantic-web-journal.net).
Gao et al. (2023), ALCE	Citation support for generated answers	On one question set, models lack complete citation support 50% of the time (arxiv.org).
Zhang and Soh (2024), EDC	Extraction and canonicalization on three benchmarks	WebNLG test: 1165 pairs (aclanthology.org). Reference triplets are non-exhaustive, so valid extra triplets are penalized (aclanthology.org).
AEVS authors (2026)	Residual hallucination after anchoring	Rates from 0.23% to 20.23% across 4 backbone models and 3 datasets, rising with dataset difficulty (www.mdpi.com).

Study (year)	What was measured	Reported result and scope
MEDAKA authors (2025)	Human review of extracted relations from drug leaflets	3,427 of 3,549 sampled relations correct (96.6%), 34 incorrect and 88 partially correct (arxiv.org).

The table has three lessons. First, labels are not ground truth: the rate at which reference sets mislabel or omit relations is large (Table 2). Second, the same system can look very different under a presence check and under entailment review, as the 65.2% versus audit result shows (arxiv.org). Third, exact-match scoring against references can understate quality. A study of language-model relation extraction found that in a manual re-judgment about half of apparent false positives on one dataset were valid, and it concluded that strict exact matching is pessimistic (arxiv.org). That study used GPT-3 and Flan-T5, so it does not reflect current systems.

Annotation agreement also varies. FEVER reported a Fleiss kappa of 0.6841 for claim labelling (arxiv.org). The EDC authors report a high inter-annotator agreement for their manual evaluation of self-canonicalization with two annotators (aclanthology.org).

Computing a support rate and an unsupported-claim rate

A rate is only defined with a disclosed sample, labels and denominators. The calculation below uses invented counts (Hypothetical Example) to show the arithmetic, not a measured system.

A reviewer set labels a random sample of 200 accepted-candidate edges: 164 supported, 14 contradicted, 17 insufficient and 5 inaccessible.

- **Support rate, all sampled edges:** 164 / 200 = 82.0%, Wilson 95% interval 76.1% to 86.7%.
- **Support rate, accessible edges only:** 164 / 195 = 84.1%, interval 78.3% to 88.6%. The 5 inaccessible edges are excluded and disclosed.
- **Unsupported-claim rate:** (14 contradicted + 17 insufficient) / 195 = 15.9%, interval 11.4% to 21.7%.
- **Inaccessible rate:** 5 / 200 = 2.5%, reported separately, because an unreadable source is a coverage problem, not evidence of falsity.

The Wilson interval is used because a US Air Force Institute of Technology review of binomial intervals reports that it performs well and is generally viewed favorably in the reviewed literature (www.afit.edu). If the sample is stratified (for example by source type or by predicate), compute rates within strata and weight them, and state the weights. FActScore offers a precedent for the unit: the atomic fact, scored as supported or not by a chosen knowledge source.

Reviewer disagreement should be reported next to the rates. In a further invented example (Hypothetical Example), suppose 60 edges were double-reviewed and the two reviewers both marked 40 supported and both marked 12 not supported, while they split on 8 (5 and 3). Percent agreement is 52 / 60 = 86.7%, but Cohen's kappa is 0.66 once chance agreement of 60.8% is removed. The clinical-statistics literature advises calculating both percent agreement and kappa (pmc.ncbi.nlm.nih.gov), gives a worked example in which percent agreement of 0.94 corresponds to a kappa of 0.85 (pmc.ncbi.nlm.nih.gov), and suggests that any kappa below 0.60 indicates inadequate agreement (pmc.ncbi.nlm.nih.gov), with at least 30 comparisons as a heuristic sample size (pmc.ncbi.nlm.nih.gov). Where there are more than two reviewers or missing ratings,

Krippendorff's alpha is an option; one methods paper states that bootstrapping is the only suitable confidence interval approach for it (www.ncbi.nlm.nih.gov), and cautions that reliability estimates depend strongly on category prevalence (www.ncbi.nlm.nih.gov). With rare contradictions, a high percent agreement can coexist with a low kappa.

Frequently Asked Questions (FAQs)

What is knowledge graph provenance at the source-span level?

It is a locator and verbatim text identifying the passage offered as support for an edge, tied to a document version. The anchor-based framework formalizes the constraint: if every element of a triplet can be traced back to a specific span, hallucinations become detectable (www.mdpi.com). Its authors name the lack of provenance as the root cause of unsupported triplets (www.mdpi.com).

How can LLM-extracted knowledge graph relationships be verified?

Run four checks in order: a deterministic check that the quoted text exists at the stated offsets, an interpretability check on the referents, an entailment judgment by a trained reviewer or a calibrated entailment model, and a counterevidence scan. The TRIAGE preprint proposes a similar grounding signal, flagging triples with low entailment and high contradiction probability (arxiv.org).

How do you ground citations for extracted triples?

Store the span with the edge and judge the span against the accepted claim, not against the entities. A span that merely mentions the entities fails most of the fixture cases in Table 1.

How do you track knowledge graph edge provenance?

Per edge, not per document. Existing tools record chunk, document or text-unit references (neo4j.com) (microsoft.github.io); add a character-level locator and a document version to make the reference reviewable.

What is evidence span attribution in relation extraction?

It is the practice of recording the evidence for a label. SciFact's rationale sentences, described above, are one example. A coverage measure (the fraction of triples with an evidence span available) is not an entailment test.

What makes a knowledge graph auditable?

Any accepted claim can be traced to a span, a document version, an observation date and a reviewer, and unknowns are recorded with a reason. The provenance ontology provides terms for derivation and attribution but does not itself check a passage (www.w3.org).

How common are hallucinated edges?

Rates depend on definitions. One framework measured 0.23% to 20.23% across models and datasets (www.mdpi.com); another benchmark's hallucination metrics are lexical presence checks. In the anchor-based study, more than 95% of detections were relations unsupported by the text between correctly grounded

entities. Presence of the entities is therefore a weak signal.

How accurate are GraphRAG citations?

The documentation pages and paper read for this report describe the mechanism and do not report a measured edge-level citation accuracy. Outputs reference text units (microsoft.github.io), the claim table includes a short text string and an LLM-derived assessment of correctness that is not a verified entailment result (microsoft.github.io). Measuring citation accuracy would need the labeled sample described above.

Implications and Future Directions

For builders. Treat verification as a gate with abstention, not as an average. The studies above suggest that precision gains cost coverage (ceur-ws.org), so the policy question is which unsupported edges a decision graph can tolerate, and the answer should be set per use, not per tool.

For researchers. Three gaps are open. No study reports reviewer time per edge or the cost of the two-stage procedure. No benchmark combines negation, modality, attribution, time and coreference at the edge level, so the fixture here needs independent validation. And ProVe's own authors note that it does not use qualifiers and that detecting refuting sources is hard (semantic-web-journal.net), which are the same difficulties as fixture cases 2 and 5. ProVe's evaluation set contains 416 pairs from Wikidata dumps of March 2022 (semantic-web-journal.net), so it supports claims about that sample and not universal correctness. Its arXiv version reports different overall figures in its introduction (75% accuracy and 68.1% F1-macro) than the revised journal text (arxiv.org), a reminder to cite the version.

For evidence reviewers. Insist on denominators. A support rate without the sample size, the labels, the treatment of inaccessible sources and the agreement statistic is not a measurement.

A note on tools. Readers who need a graph database, an extraction framework or a retrieval pipeline should choose those directly; the method in this report is independent of them. Reliable Context, the publisher of this report, offers a system of record for organizations that fuse private and public information. According to the publisher, its observable properties are one canonical record per real thing, a verbatim cited source and an observation date behind every value, unknowns declared as not found, blocked or not applicable with a reason, and reproducible past states, and it runs inside the customer's perimeter. It is not a vector database, data catalog or chatbot. This report does not describe its internals, has not independently verified the properties above, and uses no first-party results as evidence.

Open questions. Does span-level review change downstream decisions measurably? What is the lowest span granularity at which reviewers agree? How should graphs represent a claim supported by one source and contradicted by another? RDF 1.2 reifiers provide a place to attach per-source evidence (www.w3.org), but no study evaluates reviewers working with them.

Conclusion

A knowledge graph edge may enter a decision-support graph when a reviewer can open a located, verbatim span in a known document version and confirm that it entails the same qualified claim as the edge, after extraction, referent resolution and canonicalization. Where it does not, the edge is corrected, held as

insufficient, rejected as contradicted, or recorded as inaccessible. The published literature supplies the pieces: provenance vocabularies for the chain, span locators for the text, attribution and fact-verification protocols for the judgment, and error-analysis studies for the failure modes. What it does not supply is a validated end-to-end edge test, so any team adopting this procedure should publish its fixture, its sample, its denominators and its agreement statistics, and treat the first results as a measurement of the procedure itself.

Source links

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

1. https://aclanthology.org/D19-6607.pdf
2. https://www.mdpi.com/2073-431X/15/3/178
3. https://www.w3.org/TR/prov-o
4. https://semantic-web-journal.net/system/files/swj3467.pdf
5. https://www.semantic-web-journal.net/content/prove-pipeline-automated-provenance-verification-knowledge-graphs-against-textual-sources
6. https://arxiv.org/pdf/2112.12870
7. https://www.afit.edu/STAT/statcoe_files/12_Binomial%20proportion%20intervals%20DRAFT%20-%20PA%20copy%281%29.pdf
8. https://www.w3.org/TR/rdf12-concepts
9. https://neo4j.com/blog/knowledge-graph/how-to-build-knowledge-graph
10. https://www.w3.org/TR/rdf11-concepts
11. https://www.w3.org/TR/annotation-model
12. https://arxiv.org/pdf/2404.16130
13. https://aclanthology.org/2024.emnlp-main.548
14. https://aclanthology.org/2024.emnlp-main.548.pdf
15. https://aclanthology.org/D19-6607
16. https://neo4j.com/docs/neo4j-graphrag-python/current/user_guide_kg_builder.html
17. https://reference.langchain.com/python/langchain-neo4j/graphs/neo4j_graph/Neo4jGraph/add_graph_documents
18. https://microsoft.github.io/graphrag/index/outputs
19. https://microsoft.github.io/graphrag/index/default_dataflow
20. https://developers.llamaindex.ai/python/framework/module_guides/indexing/lpg_index_guide
21. https://docs.claude.com/en/docs/build-with-claude/citations
22. https://www.wikidata.org/wiki/Help:Sources
23. https://arxiv.org/pdf/2308.02357
24. https://arxiv.org/pdf/2602.11886
25. https://arxiv.org/pdf/2305.14627
26. https://arxiv.org/pdf/1803.05355
27. https://arxiv.org/pdf/2004.14974

28. https://arxiv.org/pdf/2305.14251
29. https://www.w3.org/TR/prov-dm
30. https://www.mediawiki.org/wiki/Wikibase/DataModel
31. https://pmc.ncbi.nlm.nih.gov/articles/PMC2586758
32. https://aclanthology.org/W10-3001.pdf
33. https://www.cs.brandeis.edu/~roser/pubs/sauriPustejovsky_lrec08_2.pdf
34. https://aclanthology.org/L08-1093.pdf
35. https://dlab.epfl.ch/people/west/pub/Vaucher-Spitz-Catasta-West_WSDM-21.pdf
36. https://www.wikidata.org/wiki/Help:Qualifiers
37. https://www.wikidata.org/wiki/Help:Ranking
38. https://www.wikidata.org/wiki/Help:Evolving_knowledge
39. https://timeml.github.io/site/publications/timeMLpubs/IWCS-v4.pdf
40. https://arxiv.org/pdf/1906.06127
41. https://aclanthology.org/Q15-1023.pdf
42. https://ceur-ws.org/Vol-4246/llms4kgoe-5.pdf
43. https://arxiv.org/pdf/2004.14855
44. https://arxiv.org/pdf/2104.08398
45. https://arxiv.org/html/2205.12696v3
46. https://arxiv.org/pdf/2509.26128
47. https://arxiv.org/pdf/2305.05003
48. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
49. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4974794
50. https://arxiv.org/pdf/2607.03447
51. https://arxiv.org/pdf/2210.14846

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.