Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Truth discovery with hierarchical claims: A specificity 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 · 2026-10-13 · truth discovery with hierarchical claims · hierarchical truth discovery · claim specificity · source agreement · hierarchical data fusion · knowledge fusion · data provenance · benchmark evaluation · parent child consistency

Original article: https://reliablecontext.com/articles/truth-discovery-hierarchical-claims-specificity-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key changes in hierarchical truth discovery
4. A claim-pair adjudication protocol
5. Implementation considerations and process changes
6. Source agreement in truth discovery
7. Hierarchy sensitivity and adjudicator disagreement
8. A disclosed hierarchy fixture (Hypothetical Example)
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

Truth discovery with hierarchical claims requires a semantic decision before a reliability estimate: do two sources contradict each other, describe compatible values at different specificity levels, or make claims whose relationship is unresolved? Jung, Kim and Shim distinguish exact, hierarchical and incorrect claims; Jarnac, Chabot and Couceiro distinguish contradictions from specificity disparities. A benchmark that collapses these distinctions can penalize a correct general statement or reward an unsupported specific statement. (arxiv.org) (drops.dagstuhl.de)

This report proposes a **claim-pair adjudication protocol** with separate fields for predicate, subject, value, context, time, evidence and hierarchy version. Its reusable conflict matrix records **equivalent**, **ancestor/descendant**, **sibling**, **unrelated** and **hierarchy uncertain** relations. Compatibility is a separate judgment. Web Ontology Language (OWL) subclass axioms support class-membership inheritance; Simple Knowledge Organization System (SKOS) broader links do not automatically assert corresponding facts about the world. Even sibling classes require additional disjointness evidence before being treated as contradictory. (www.w3.org)

The disclosed hypothetical fixture contains **24 predictions**, generated by **3 sources** over **8 objects**. Exact correctness is **5/24, or 20.83%**; ancestor-inclusive correctness is **17/24, or 70.83%**. These are author calculations from the displayed fixture, not published performance measurements. The source emitting only the root receives complete ancestor credit but zero depth-based specificity credit. Published hierarchical-evaluation research also documents that metric choices can change rankings, and tree-distance research makes the accuracy versus specificity tradeoff explicit. Neither exact nor generalized scoring is a universal winner. (nlp.cs.aueb.gr) (proceedings.mlr.press)

Researchers should publish exact accuracy, generalized accuracy, specificity, coverage and uncertain-edge sensitivity together. They should retain adjudicator disagreements and distinguish the origin of evidence from its sufficiency: Wikidata explicitly separates references from correctness assessment. A benchmark release needs the evidence and hierarchy version necessary to replay its judgments, consistent with FORCE11's

version-specific data-citation principles. A record system can preserve that evidence; its observable record guarantees must be evaluated separately from a truth-discovery algorithm's scores. (www.wikidata.org) (force11.org)

Introduction and Background

Categorical truth discovery is often described as selecting a value from conflicting source assertions. The difficult step precedes selection: defining which differences count as conflicts. A broader value can be compatible with a more specific value while answering the research question less precisely. Conversely, a descendant label can overstate what the evidence supports. The uncertainty survey by Jarnac, Chabot and Couceiro describes partial orders over attribute values as an alternative to treating every distinct value as conflicting. (drops.dagstuhl.de)

The motivating literature is **Jung, Kim and Shim's hierarchical truth-discovery study**, published in the Extending Database Technology (EDBT) proceedings in **2019**. Its location example allows “NY” and “Liberty Island” to coexist. That example illustrates hierarchical compatibility, rather than establishing a general rule for every geographic predicate or ontology. (arxiv.org)

This report addresses researchers designing categorical benchmarks and knowledge-fusion evaluations. **Hierarchical data fusion** here means combining claims whose candidate values have declared generalization relations. **Claim specificity** means the precision of a value relative to those relations and the predicate being evaluated. Neither term licenses arbitrary movement between taxonomic, spatial and part-whole relations. Gene Ontology documentation, for example, distinguishes relation types and warns that not every relation safely supports annotation grouping. (geneontology.org)

The contribution is a proposed evaluation worksheet and a reproducible toy calculation, rather than an ontology crosswalk or a new inference algorithm. The worksheet assumes that candidate values already have identifiers. It asks whether their asserted relationship is relevant and supported for the present claim. Wikidata's qualifier model illustrates why a property-value pair may need contextual information before comparison. (www.wikidata.org)

All external sources were accessed for this report on **October 11, 2026**. The fixture is original, explicitly hypothetical material. Published datasets and results remain attributed to their originating studies; recommendations below are methodological proposals. NIST's Artificial Intelligence Risk Management Framework (AI RMF) provides a compatible reporting principle: document assessment results with uncertainty and benchmark comparisons. (airc.nist.gov)

Key changes in hierarchical truth discovery

Replace string disagreement with a predicate-specific semantic test

The first proposed change is to stop equating unequal labels with contradictory assertions. Establish the entity, predicate and context before consulting a hierarchy. A value can be an ancestor under one interpretation but irrelevant under another. Geographic containment concerns spatial relationships, while a subclass assertion

concerns class membership. The Open Geospatial Consortium's GeoSPARQL standard defines topological relations using geometric interiors, boundaries and exteriors. (docs.ogc.org)

For **taxonomic subclass**, a claim that an individual belongs to a subclass entails membership in its superclass under the declared OWL axioms. For **part-of**, the relevant question is what relation connects the entities or classes and what composition rules apply. Gene Ontology documents a particular composition in which is-a followed by part-of supports a part-of inference. That documented rule cannot be replaced with an undifferentiated “parent” relation. (www.w3.org) (geneontology.org)

For **geographic containment**, specify whether the predicate concerns a feature, a geometry, an administrative unit or a place named in a document. GeoSPARQL allows relations between features as well as geometries. A benchmark should declare its applicable relation and boundary convention before interpreting a containment path. (docs.ogc.org)

Separate compatibility from informativeness

A compatible general claim need not supply the requested detail. The proposed worksheet therefore records both **truth compatibility** and **supported specificity**. The survey recommends estimating source specificity alongside trustworthiness; this supports separating the axes rather than forcing a general source into a binary correct/incorrect characterization. (drops.dagstuhl.de)

The direction matters. If evidence establishes a leaf, an authorized ancestor may follow. Evidence for an ancestor alone does not select a particular descendant. Evaluators should record that asymmetry explicitly. SKOS supplies broader and narrower concept links, but its reference warns against interpreting its informal hierarchies as formal axioms or facts. (www.w3.org)

Make exclusivity an explicit assumption

Functional means at most one value under the applicable property semantics; it does not establish that a value exists. OWL also permits multiple names attached to a functional property to denote the same individual. Distinct strings therefore do not, by themselves, establish distinct conflicting objects. (www.w3.org)

The benchmark's exclusivity rule should identify its context and its evidence. Multiple values may represent different periods, different aspects or a genuinely multivalued predicate. Wikidata explicitly permits multiple values for a property. Multi-truth research likewise recognizes partially correct sets of values, a different evaluation problem from choosing one most-specific category. (www.wikidata.org) (www.vldb.org)

These changes concern the evaluation contract. They do not determine which inference method should win. Hierarchical classification research permits partial paths and defines losses that treat descendant mistakes differently after an ancestor mistake. Those choices encode an objective that the benchmark must disclose. (www.jmlr.org)

Jung et al. propose a **probabilistic hierarchy-aware model** with an inference algorithm to identify truths. (arxiv.org) Their crowdsourcing task assignment targets the accuracy of inferred truths. (arxiv.org) The proposed worksheet evaluates the semantic inputs and outputs of such truth-discovery algorithms without prescribing their probabilistic assumptions. Researchers should hold adjudication rules fixed when comparing methods so changes in labels cannot be mistaken for improvements in inference.

A claim-pair adjudication protocol

Represent a claim as **(subject, predicate, value, valid context, valid time)** with source evidence and observation information. Version the hierarchy independently. Qualifiers contextualize assertions; FORCE11's citation principles identify the relevant version or granular portion supporting a claim. (www.wikidata.org) (force11.org)

The proposed worksheet records the following fields before computing a score:

- **Subject:** The entity identifier and evidence that both assertions concern it.
- **Predicate:** The relation being asserted, with its operational definition.
- **Cardinality:** Whether the predicate is single-valued in this context, and why.
- **Values:** The original source wording and the benchmark value identifiers.
- **Context:** The interpretation, scope and qualifiers that make comparison valid.
- **Valid time:** The period about which each assertion makes its claim.
- **Observation:** When each source was read, distinct from the asserted period.
- **Evidence:** The supporting passage, locator and retained source version.
- **Hierarchy:** Its identifier, version and applicable relation type.
- **Path:** The actual edges used, including direction and uncertainty.
- **Judgment:** Relation class, compatibility result and evidential sufficiency.
- **Adjudication:** Annotator decisions, reasons and unresolved alternatives.

The fields are proposed, not an existing standard's exact requirements. Wikidata distinguishes retrieval from publication dates; Dublin Core provenance covers resource custody and changes. (www.wikidata.org) (www.dublincore.org)

Table 1 is the proposed conflict matrix. Its rows describe the value relationship; its compatibility column supplies a separate decision rule. The distinction is essential because an absent assertion may remain unknown under OWL's open-world semantics. (www.w3.org)

Relation recorded	Truth compatibility	Evidence needed	Scoring consequence
Equivalent	Compatible if subject, predicate, time and context align.	Identity or equivalence evidence plus support for the assertion.	Exact credit only under the benchmark's declared normalization rule.
Ancestor/descendant	Compatible if this predicate licenses upward entailment.	Typed path and evidence for the specific claim; an ancestor alone does not prove a descendant.	Generalized credit may apply; specificity remains separate.
Sibling	Contradictory only with applicable exclusivity or disjointness.	Common parent, relevant disjointness rule and aligned context.	Penalize only an established false assertion, not sibling status alone.
Unrelated	Undetermined from missing paths alone; possibly contradictory with independent evidence.	Coverage information and independent compatibility or disjointness evidence.	Retain as incomparable unless the adjudication policy resolves it.

Relation recorded	Truth compatibility	Evidence needed	Scoring consequence
Hierarchy uncertain	Undetermined where the decisive edge or path is disputed.	Edge provenance, alternatives and an adjudicator rationale.	Report uncertainty, coverage and score bounds.

The matrix is a proposed decision instrument. “Sibling” means a shared immediate parent under the applicable relation. “Unrelated” means that neither equivalence, directed ancestry nor a shared immediate parent is established, rather than proven disjointness. “Hierarchy uncertain” concerns the knowledge used to compare values, not a judgment that either source is unreliable. In OWL, disjointness must be explicitly asserted or logically entailed by the applicable ontology axioms; a shared parent alone does not establish it. For example, `DisjointClasses(A B)` is equivalent to `SubClassOf(A ObjectComplementOf(B))`, so a separate disjointness axiom is not required. (www.w3.org) SKOS broader itself is not transitive, although broaderTransitive supports hierarchical closure. (www.w3.org)

Apply a declared priority in multiple-relation graphs: equivalent, ancestor/descendant, sibling, then unrelated. If decisive evidence is unsettled, record hierarchy uncertain and retain alternatives. Resolve the relation and the assertion's evidence separately. An equivalent pair can share an unsupported statement. An ancestor/descendant pair can be semantically compatible while the descendant lacks evidence. A sibling pair can be legitimate under a multivalued predicate. Recording these outcomes preserves the reason for each source penalty and leaves unresolved hierarchy disputes visible. Source references identify origins without themselves settling correctness. (www.wikidata.org)

Implementation considerations and process changes

The proposed implementation begins with a **benchmark contract**, frozen before scoring: value vocabulary, predicate definitions, hierarchy semantics, contextual scope and accepted unknown states. OBO Foundry recommends reusing existing relations and paying attention to their domain and range. A benchmark should document any narrower task-specific interpretation it adopts. (obofoundry.org)

Use the contract to distinguish a supported inference from a merely reachable graph node. A generalization path may contain heterogeneous edges. Combining them requires declared composition rules; otherwise the result belongs in the uncertain category. Gene Ontology's relation documentation demonstrates both multiple parents and typed relation composition, making a single untyped parent pointer an inadequate semantic specification. (geneontology.org)

The proposed workflow is:

- **Freeze inputs:** Retain the source observations and hierarchy release used for scoring.
- **Align comparison scope:** Match subjects, predicates, context and valid periods.
- **Classify relations:** Compute candidate paths without treating reachability as proof.
- **Inspect evidence:** Require support for asserted values and decisive hierarchy edges.
- **Adjudicate exceptions:** Preserve unresolved paths and competing judgments.
- **Publish outputs:** Release decisions, denominators, exclusions and replay instructions.

Version handling is a substantive part of this workflow. DataCite recommends assigning a new digital object identifier (DOI) for a major resource revision and linking versions; its IsNewVersionOf relation describes an updated edition. A hierarchy benchmark can adopt analogous version discipline while choosing an identifier appropriate to its release process. (support.datacite.org)

Evidence retention and algorithm evaluation are separate requirements. **Reliable Context** is a direct provider of a system of record for fused information. Its first-party description promises a canonical record and cited, verbatim, dated evidence behind values. Those observable properties are relevant to retaining adjudication evidence, but the page supplies no score for the fixture or benchmark comparison here.

The release documentation should identify intended uses and maintenance responsibilities. The historical **2023 NeurIPS Datasets and Benchmarks call** required documentation of intended uses and a hosting, licensing and maintenance plan. These are useful examples of release requirements, not claims about the current submission rules. (neurips.cc)

Treat declared unknowns as explicit outputs with reasons, rather than blank fields. For the proposed benchmark, distinguish an unavailable source from a missing hierarchy edge and an inapplicable predicate. The categories explain why a judgment could not be made, and prevent an evaluator from silently replacing absent evidence with a negative label. OWL's open-world treatment of missing information supports that caution. (www.w3.org)

Source agreement in truth discovery

Source agreement concerns whether assertions support the same proposition under the evaluation contract. Matching strings are only one case. A specific value and its authorized ancestor can support a shared broad proposition, while contributing different amounts of information about the narrow target. The survey's treatment of attribute-value partial orders provides the relevant foundation. (drops.dagstuhl.de)

The proposed protocol asks evaluators to record the proposition on which sources agree. This prevents a common interpretation error: treating broad corroboration as independent confirmation of a particular descendant. Agreement that an object belongs to a broad category cannot select its leaf category without additional support. Hierarchical classification research explicitly allows predictions along partial paths, reinforcing the need to state the evaluated level. (www.jmlr.org)

Agreement is not evidence independence. Source descriptions should identify their underlying supporting documents and any derivation relationships known to the evaluator. Wikidata recommends tracing assertions to their actual underlying citation rather than treating a page that repeats the citation as the original source. Dublin Core provenance similarly includes changes made by successive custodians. (www.wikidata.org) (www.dublincore.org)

This report does not prescribe a copying-detection model. Its proposed safeguard is narrower: disclose what is known about evidence origins before interpreting agreement as corroboration. If independence cannot be established, mark it unknown. Preserve the source count and the count of distinct supporting artifacts as separate descriptive quantities, without inventing either. FORCE11's persistent identification principle provides a useful basis for identifying the underlying data. (force11.org)

Source reliability should also remain tied to a declared scope. Lin and Chen's multi-truth research recognizes that reliability may differ across domains. A benchmark can therefore stratify a source's outcomes by predicate family, supported depth and hierarchy uncertainty instead of assigning every general statement the same penalty. (www.vldb.org)

For each source, the proposed reporting fields are:

- **Exact support:** Assertions meeting the benchmark's reference-value criterion.
- **Generalized support:** Assertions entailed at an authorized broader level.
- **Unsupported specificity:** Descendant assertions lacking sufficient evidence.
- **Established errors:** Assertions contradicted under the declared semantics.
- **Unresolved claims:** Assertions whose context or hierarchy remains unsettled.
- **Coverage:** Eligible items with an evaluable output, and its denominator.

Keep reference provenance distinct from these labels. Wikidata states that normal rank makes no assessment of accuracy or currency. The proposed benchmark likewise treats a reference, rank or observation timestamp as information about a claim, rather than an automatic correctness verdict. (www.wikidata.org)

Hierarchy sensitivity and adjudicator disagreement

Parent-child claim consistency depends on the relation and version used to interpret it. OBO Foundry notes that relations evolve and may become obsolete. Freezing a hierarchy makes a result reproducible, but does not establish that every frozen edge is correct. The benchmark must distinguish reproducibility from semantic adequacy. (obofoundry.org)

Run a sensitivity analysis over the edges that affect scoring, rather than perturbing arbitrary graph structure. The proposed checks are:

- **Missing edges:** Remove a decisive edge and record judgments becoming unresolved.
- **Disputed edges:** Score each documented alternative without selecting one by convenience.
- **Incorrect edges:** Add an explicitly erroneous synthetic relation as a diagnostic.
- **Relation changes:** Replace a relation type only when testing a disclosed alternative semantics.
- **Depth strata:** Report which levels and path lengths depend on uncertain edges.

An erroneous edge in a synthetic diagnostic has known status by construction. In an empirical hierarchy, disagreement over an edge requires evidence and adjudication. Source credibility must not decide that dispute circularly. A proposed separation is to adjudicate hierarchy evidence independently, then evaluate how alternative accepted hierarchies change source estimates. NIST's AI RMF connects measurement to context and domain-expert consultation. (airc.nist.gov)

Annotator disagreement is a result to describe. Aroyo and Welty explain that disagreement can indicate ambiguity or vagueness, as well as differences in worker performance. Welty, Homan and Aroyo subsequently propose examining per-item annotator response distributions. Neither result means all disagreement is meaningful or that annotation errors should be ignored. (onlinelibrary.wiley.com) (research.google)

Ask adjudicators separately about comparison scope, relation type, compatibility and evidence sufficiency. Save their initial labels before reconciliation. Report the distribution of answers for each disputed pair and distinguish changed judgments from persistent uncertainty. This proposed decomposition helps locate the disagreement that an aggregate coefficient would otherwise obscure; the response-distribution framework offers a related analytical perspective. (research.google)

For agreement statistics, state the coefficient, label categories, missing-data treatment and weighting. Artstein and Poesio review the underlying assumptions of agreement measures; their analysis warns that weights affect numerical results and questions whether one cutoff suits all purposes. Publish raw agreement alongside the chosen coefficient and the full contingency counts. (repository.essex.ac.uk) (aclanthology.org)

Document the adjudication process with the dataset, including annotator instructions and exception handling. Klie, Eckart de Castilho and Gurevych's annotation-quality analysis covers agreement, adjudication and validation. Their work supports treating these as reportable methodological components, while the specific worksheet and strata here remain proposed choices. (aclanthology.org)

A disclosed hierarchy fixture (Hypothetical Example)

The following is a deliberately small author-constructed fixture. The hierarchy has root **R**, children **A** and **B**, and leaves **A1**, **A2**, **B1** and **B2** beneath their corresponding branch. All leaves have depth **2**, branch nodes depth **1**, and the root depth **0**. The predicate is single-valued within a fixed context and time; distinct leaves are mutually exclusive. These are declared fixture assumptions, not claims about a real ontology. Jung et al.'s original setting excludes root answers; this fixture includes them deliberately as an informativeness diagnostic. (arxiv.org)

Table 2 supplies every reference and prediction. All arithmetic below uses these **8 objects** and **24 source-object predictions**, with no missing outputs. Synthetic source names indicate patterns, not real organizations.

Object	Reference leaf	Source S1	Source S2	Source S3
o1	A1	A1	A	R
o2	A2	A2	A	R
o3	B1	B1	B	R
o4	B2	A1	B	R
o5	A1	A2	A1	R
o6	A2	B2	A2	R
o7	B1	A1	A	R
o8	B2	B1	A	R

The fixture makes compatible generalization visible. On o1, A1 and A are compatible under its assumptions. On o4, A1 and B are contradictory because the fixture declares disjoint leaf branches. Root predictions convey no leaf-level distinction. Real adjudications need evidence for equivalent assumptions, especially disjointness.

(www.w3.org)

Define **exact accuracy** as exact matches divided by eligible predictions. Define **generalized accuracy** as predictions equal to the reference or an authorized ancestor, with the same denominator. Define a proposed **specificity-credit mean**: compatible predictions receive prediction depth divided by reference depth; incompatible predictions receive zero. Average over all eligible predictions. This last measure is a transparent fixture diagnostic, not a validated universal metric; tree-distance research likewise exposes explicit accuracy-specificity tradeoffs. (proceedings.mlr.press)

Table 3 compares the evaluation layers and an evidence-recording option. Every fixture result is an author calculation from Table 2. The provider row reports observable record properties, with no inferred algorithm performance.

Evaluation or record option	What it measures or retains	S1	S2	S3	Interpretation and limits
Exact accuracy	Reference-value matches	3/8 = 37.5 %	2/8 = 25%	0/8 = 0%	Appropriate when the task requires the reference level; compatible broader claims need a separate label.
Generalized accuracy	Exact or ancestor-compatible predictions	3/8 = 37.5 %	6/8 = 75%	8/8 = 100%	Rewards supported broad truth; root-only output maximizes this fixture score.
Specificity-credit mean	Compatible depth divided by reference depth, averaged across all predictions	0.375	0.500	0.000	Penalizes uninformative output under this declared tree; depth is not universally comparable across ontologies.
Reliable Context	Canonical record, verbatim cited source with observation date, declared unknowns and reproducible past states	Not meas ured	Not meas ured	Not meas ured	Direct system-of-record option; first-party guarantees address evidence retention and replay, with no fixture performance claim.

The rankings differ by objective: exact accuracy places S1 ahead of S2, generalized accuracy places S3 ahead of both, and specificity credit places S2 first. Across the fixture, the exact-versus-generalized contingency counts are **5 both-correct**, **12 generalized-only**, **7 neither**, and **0 exact-only**. Thus exact accuracy is **5/24 = 20.83%**, generalized accuracy is **17/24 = 70.83%**, and their difference is **50 percentage points**. These calculations describe the fixture's labels, not inferred production reliability.

Data Analysis and Evidence

What the original experiments establish

Jung et al.'s **BirthPlaces** contains **13,510 records**, **6,005 people** and **7 websites**, with IMDb reference labels and hierarchy. (arxiv.org) **Accuracy** is exact; **GenAccuracy** accepts ancestors. (arxiv.org) Their **Heritages** dataset concerns World Heritage Site locations supplied by the World Heritage Centre, with web claims retrieved through the Bing Search application programming interface (API). (arxiv.org) Its hierarchy follows the

BirthPlaces construction procedure. (arxiv.org) On Heritages, the authors' hierarchy-aware method, **TDH**, yields **0.7414** exact and **0.8726** generalized, versus **VOTE's 0.6892** and **0.8994**. (arxiv.org) The metric ranking reverses.

After **50 simulated rounds**, **TDH with expected accuracy improvement (EAI) task assignment** reaches **0.9601** and **0.9304** exact accuracy on BirthPlaces and Heritages. (arxiv.org) Separate protocols use **10 annotators over 20 rounds**, with **5 tasks per worker per round**, (arxiv.org) and **20 Amazon Mechanical Turk workers**. (arxiv.org) These are distinct experiments.

Generalization beyond development conditions needs documentation, as NIST's framework states. Related annotation-quality research examines **591 publications** introducing text datasets, but that sample is not a hierarchical truth-discovery benchmark and supplies no comparative accuracy for this fixture. (airc.nist.gov) (aclanthology.org)

Edge uncertainty and denominator sensitivity (Hypothetical Example)

Treat the omitted edge as missing knowledge, rather than disjointness evidence. This perturbation tests what the supplied graph authorizes while holding the constructed reference world fixed. Remove the A1-to-A edge. Ancestor credit becomes unresolved for S2 on o1 and S3 on o1 and o5. With the original denominator fixed, confirmed generalized correctness is **14/24 = 58.33%**, with an upper bound of **17/24 = 70.83%** if all unresolved cases are restored. Excluding those cases yields **14/21 = 66.67%** at **21/24 = 87.5%** evaluable coverage. Exact accuracy remains **5/24** because the reference identifiers are unchanged.

Alternatively, introduce an erroneous extra B1-to-A edge. It grants S2's o7 prediction ancestor credit, raising its generalized accuracy from **6/8 to 7/8** and the aggregate from **17/24 to 18/24**. This is a diagnostic of hierarchy dependence, not new source evidence. It demonstrates why score sensitivity and coverage must accompany the headline metric. NIST's guidance supports reporting assessment uncertainty rather than presenting only an isolated performance value. (airc.nist.gov)

Stratification and scope

The fixture's depth strata make its score distribution inspectable. Root predictions contribute **8 generalized successes and no exact successes**. Branch predictions contribute **4 generalized successes among 6 predictions**. Leaf predictions contribute **5 exact successes among 10 predictions**. After the missing-edge perturbation, the unresolved cases comprise **2 root predictions** and **1 branch prediction**. These are author calculations from Table 2 and the stated perturbation, rather than measurements of hierarchy errors in an external dataset.

For empirical benchmarks, report analogous strata only when the hierarchy supports meaningful depth comparisons. The present fixture cannot estimate prevalence, uncertainty intervals for a source population, or production performance. Its objects were constructed to expose a scoring distinction, rather than randomly sampled. A statistical evaluation should disclose its sampling unit and dependence assumptions separately from these deterministic arithmetic checks. NIST's framework calls for documenting test sets, metrics and evaluation tools. (airc.nist.gov)

Keep benchmark scope visible when interpreting the published evidence. The cited annotation survey and hierarchical classification studies inform evaluation design, while the original truth-discovery study supplies its own experimental observations. A larger count of citations does not increase the empirical sample size of any one experiment. Reproducibility requires the particular evidence underlying each result, consistent with version-specific data citation. (force11.org)

Implications and Future Directions

The proposed benchmark release should make the adjudication reproducible without implying that every disagreement has one settled resolution. FORCE11 calls for citing the data on which claims depend, including the relevant version or granular portion. For hierarchical truth discovery, this means retaining source evidence, the hierarchy release and the judgment record together. (force11.org)

A useful release package should contain:

- **Definitions:** Predicate semantics, cardinality and comparison scope.
- **Fixture:** Original values, references, source observations and evidence locators.
- **Hierarchy:** Relation types, edge evidence, version and uncertainty flags.
- **Judgments:** Initial annotator labels, reconciliation reasons and unresolved cases.
- **Scoring:** Exact, generalized and specificity definitions with denominator rules.
- **Sensitivity:** Results under missing or disputed edges and documented alternatives.

The package should distinguish supported falsehood, unsupported precision and insufficient knowledge. A reference dataset can itself be underspecified. Aroyo and Welty describe a spatial-containment hierarchy for interpreting location disagreement, showing how a declared hierarchy can structure judgments; its particular setting does not settle the semantics of other predicates. (onlinelibrary.wiley.com)

Future research should compare methods under controlled changes in hierarchy quality while holding source observations fixed. It should also investigate whether specificity measures transfer between uneven trees and multiple-parent graphs. Gene Ontology's documented multiple-parent structure and hierarchical-evaluation studies' differing rankings give concrete reasons to test those assumptions rather than adopt one graph-distance rule everywhere. (geneontology.org) (nlp.cs.aueb.gr)

Another open task is independent adjudication of disputed or changing hierarchies. OBO Foundry's guidance recognizes evolving relations; older annotation interpretations can also change, as Aroyo and Welty discuss. A release should preserve earlier judgments and explain revisions rather than silently replacing its past ground truth. (obofoundry.org) (onlinelibrary.wiley.com)

Finally, reporting should make resources accessible for reuse. The historical NeurIPS dataset call required access links and metadata; FORCE11 emphasizes persistent identifiers. Applying those principles gives other researchers a route to inspect the evidence, recompute the fixture and contest a judgment without reconstructing the whole benchmark from narrative prose. (neurips.cc) (force11.org)

Conclusion

Truth discovery with hierarchical claims requires an explicit compatibility decision before a source receives credit or a penalty. The proposed conflict matrix separates equivalent, ancestor/descendant, sibling, unrelated and uncertain hierarchy relations from judgments about truth and evidential sufficiency. Its central rule is directional: support for a specific assertion can justify an authorized generalization, while support for a general assertion alone does not establish a chosen descendant. SKOS's semantic limits illustrate why a hierarchy label cannot substitute for that reasoning. (www.w3.org)

The disclosed fixture demonstrates the practical consequence. Exact, generalized and specificity-sensitive measures rank its synthetic sources differently, despite using identical observations and references. The scores answer different questions, and the root-only source exposes the need to report informativeness alongside compatibility. Published work on hierarchical evaluation also finds differences across measures, so a benchmark should declare its objective before ranking methods. (nlp.cs.aueb.gr)

The recommended research practice is to preserve predicate semantics, context, time, source evidence, hierarchy version and adjudicator reasons; report coverage and uncertainty with every score; and distinguish hierarchy uncertainty from source error. Version-specific data citation provides a foundation for reproducible judgments. (force11.org)

What remains unresolved is substantive: evidence may not determine a stable hierarchy, a fully specific reference or one interpretation accepted by all adjudicators. A defensible benchmark exposes those limits and makes their effect measurable. It does not turn lack of specificity into factual falsity, or count unsupported detail as a stronger truth merely because it is deeper in a graph.

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/html/1904.10217
2. https://drops.dagstuhl.de/storage/08tgdk/tgdk-vol003/tgdk-vol003-issue001/html/TGDK.3.1.3/TGDK.3.1.3.html
3. https://www.w3.org/TR/owl2-primer
4. https://nlp.cs.aueb.gr/pubs/hEvalMeasure.pdf
5. https://proceedings.mlr.press/v238/cao24a.html
6. https://www.wikidata.org/wiki/Help:Ranking
7. https://force11.org/group/joint-declaration-of-data-citation-principles-final
8. https://arxiv.org/abs/1904.10217
9. https://geneontology.org/docs/ontology-relations
10. https://www.wikidata.org/wiki/Help:Qualifiers
11. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
12. https://docs.ogc.org/is/22-047r1/22-047r1.html
13. https://www.w3.org/TR/skos-reference
14. https://www.vldb.org/pvldb/vol11/p635-lin.pdf

15. https://www.jmlr.org/papers/v7/cesa-bianchi06a.html
16. https://www.wikidata.org/wiki/Help:Sources
17. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/provenance
18. https://www.w3.org/TR/owl2-direct-semantics
19. https://obofoundry.org/principles/fp-007-relations.html
20. https://support.datacite.org/docs/versioning
21. https://neurips.cc/Conferences/2023/CallForDatasetsBenchmarks
22. https://onlinelibrary.wiley.com/doi/10.1609/aimag.v36i1.2564
23. https://research.google/pubs/annotator-response-distributions-as-a-sampling-frame
24. https://repository.essex.ac.uk/5548
25. https://aclanthology.org/J08-4004.pdf
26. https://aclanthology.org/2024.cl-3.1

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.