

PROV-CRED: Rated Provenance for Knowledge Graphs

Mary Ann Tan^{1,2}, Christian Nennemann³, Genet Asefa Gesese^{1,2} and Harald Sack^{1,2}

¹FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany

²Karlsruhe Institute of Technology, Institute AIFB, Germany

³Independent Researcher

Abstract

AI systems that reason over knowledge graphs (KGs) should treat a triple’s reliability and credibility as first-class, queryable properties, weighting, filtering, or flagging facts by how reliably they were produced rather than consuming every triple as equally authoritative. Current KG-grounded reasoning falls short of this. Triples derived from deterministic schema rules, heuristic matching, and model prediction are treated as equivalent in reliability and credibility, though their accuracy and failure modes differ. A hallucinated relation or mislinked entity deposited alongside schema-mapped facts enters downstream retrieval without distinction; errors recirculate as authoritative ground truth, compounding with each pass. PROV-O records that derivation occurred but not the enrichment method, producing model, or benchmark performance, a gap that also applies to agentic systems maintaining graph-backed memory. As a step toward the vision, PROV-CRED addresses the resulting reliability opacity problem as a minimal PROV-O extension, recording enrichment method, model benchmark performance, and institutional accreditation as queryable PROV-O properties of each triple. Calibration, self-evaluation bias, annotation granularity, retroactive provenance, domain mismatch, and, most centrally, the governance question of who may accredit a model’s outputs remain open, together forming the paper’s research agenda toward AI systems that can act on the reliability and credibility of every triple they consume.

Keywords

knowledge graph provenance, knowledge graph enrichment, credibility annotation, reliability annotation, PROV-O extension, RDF-Star, AI accountability

1. Introduction

For AI systems that reason over knowledge graphs (KGs), a triple’s reliability and credibility should be first-class, queryable properties: facts get weighted, filtered, or flagged by how reliably they were produced, not treated as uniformly authoritative. This is the vision this paper pursues. KG-grounded reasoning implicitly assumes that triples carry uniform reliability and credibility regardless of how they were produced. As agentic AI systems combine LLM reasoning with KG-grounded memory [1], this assumption is under growing pressure. KG construction is no longer a single pipeline but a heterogeneous mixture of enrichment processes, including schema-based mapping, heuristic string matching, knowledge graph embedding (KGE) completion, and large language model (LLM) entity and relation extraction. Although these methods differ significantly in accuracy, robustness, and failure modes, their results are ultimately deposited in the same triple store. At query time, a triple mapped from a controlled vocabulary and a triple predicted by a KGE model are identical in form: both are plain RDF triples with a subject, predicate, and object.

In KG-backed retrieval and reasoning systems [2, 3], triples are retrieved without access to their method of origin; heterogeneous facts are treated as having comparable reliability. Hallucination in KG-grounded generation is a documented failure mode [4]; work on LLM–KG integration [5] improves factual reliability but does not address the source: the KG provides no basis for distinguishing a reliable triple from an unreliable one.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

✉ ann.tan@fiz-karlsruhe.de (M. A. Tan); research@nennemann.de (C. Nennemann); genet-asefa.gesese@fiz-karlsruhe.de (G. A. Gesese); harald.sack@fiz-karlsruhe.de (H. Sack)

🆔 0000-0003-3634-3550 (M. A. Tan); 0009-0004-9500-9602 (C. Nennemann); 0000-0003-3807-7145 (G. A. Gesese); 0000-0001-7069-9804 (H. Sack)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This creates a structural limitation: the KG itself does not encode the enrichment method of a triple, the model that produced it, or its benchmark performance on the relevant task, nor does it capture how reliable that generation process is expected to be under comparable conditions. While pipeline engineers may know which model was used and what score it achieved, no standard vocabulary exists to record this information, and it is typically lost once the data is written to a triple store.

We refer to this as the reliability opacity problem in KGs: heterogeneous enrichment processes produce semantically equivalent triples without standardized, queryable metadata about their generation method or empirical reliability.

As KG-grounded AI systems are deployed in Annex III high-risk domains under the EU AI Act¹, the inability to distinguish a schema-mapped triple from an LLM-predicted one is a compliance gap. Article 13 requires documentation of capabilities and limitations, Article 15 requires accuracy documentation for the intended purpose, and Annex IV §2(g) requires per-component performance metrics. The Act regulates the *AI system*, not a standalone KG, so the compliance hook activates when the KG is embedded as a component in a high-risk system. PROV-CRED operationalizes these obligations at the component level: benchmark performance is recorded in the graph alongside each produced triple, machine-readable and queryable rather than confined to a static annex.

Confidence annotation at the fact level is feasible but has not produced a reusable vocabulary: NELL [6] has maintained confidence-annotated facts since 2010, yet no general scheme emerged. Three conditions that did not hold in 2010 now make standardization viable: (1) PROV-O is widely deployed infrastructure, so an extension carries lower adoption cost than a standalone system; (2) RDF-Star [7] in RDF 1.2 provides triple-level annotation without the reification overhead that made earlier approaches impractical; and (3) demand for credibility distinction is new: LLM-KG integration pipelines such as [3, 5, 8] were not widespread when earlier annotation schemes were proposed.

To make the vision tractable, PROV-CRED extends the W3C PROV-O provenance ontology [9] with three additions.

1. An activity subclass hierarchy `HumanEnrichment`, `DeterministicEnrichment`, `HeuristicEnrichment`, `ModelEnrichment`, encodes enrichment method as the type of the generating `prov:Activity`, queryable via `prov:wasGeneratedBy`.
2. A `provcred:Model` class records model identity, version, task type, and benchmark performance; a `provcred:ReliabilityAttribution` attaches that record to each produced triple via PROV-O qualified attribution.
3. A `provcred:CredibilityAttribution` class holds the slot for institutional accreditation — the second credibility signal and the paper’s Blue Sky provocation.

`provcred:CredibilityAttribution` deliberately overstates the current scope: the vocabulary delivers one grounded dimension — method reliability grounded in benchmark evidence and inference scores when applicable, attached to each triple via `provcred:ReliabilityAttribution`. The paper’s provocation is what principled institutional accreditation would look like and how it would compose with method reliability at query time.

Sections 2–3 cover related work and vocabulary design as groundwork. Section 4 demonstrates the vocabulary on three worked examples, while open research problems and governance provocation, the paper’s central contribution, are presented in Section 5. Finally, Section 6 closes on the accreditation question.

2. Background

2.1. Probabilistic Knowledge Graphs

NELL [6] has maintained per-fact confidence scores alongside method provenance since 2010, updated continuously as extraction runs. ProBase [10] attaches confidence scores to probabilistic taxonomy assertions drawn from multiple extraction methods. YAGO2 records confidence scores for facts aggregated

¹EU AI Act, <https://artificialintelligenceact.eu/the-act/>

from multiple sources [11]. KGE models produce a ranking score for each predicted triple as a direct output of the embedding objective [12]. Post-hoc trustworthiness measurement approaches estimate triple correctness from internal KG structure and global inference [13]. PROV-CRED is orthogonal, recording production-time provenance rather than post-hoc plausibility.

No standard vocabulary has emerged for recording scores and their provenance across methods: each system uses its own representation. Scores are also not comparable across methods without calibration, meaning a cross-method mapping that places scores from embedding models, heuristic extractors, and generative models on a common scale. A KGE ranking score of 0.7452 and an LLM self-reported confidence of 0.85 are not on a common scale and cannot be directly compared without calibration across methods. Enrichment method is also not a queryable property of the triple. NELL tracks provenance at the extractor level, but the distinction between a deterministic and a probabilistic classifier, which have fundamentally different reproducibility and audit-path properties, is not represented in the graph.

2.2. Provenance Vocabularies

PROV-O [9] provides a vocabulary for recording who produced what, when, and from what source. It captures derivation lineage (`prov:wasDerivedFrom`), attribution (`prov:wasAttributedTo`), and activity descriptions (`prov:Activity`), but not the enrichment *method*, that is, whether a triple was produced by a deterministic rule, a heuristic, or a model, or the producing agent’s performance on the specific task.

Nanopublications [14] provide provenance at the individual assertion level and have been applied in scientific knowledge graphs to associate claims with their supporting evidence, but do not define enrichment method semantics or link assertion provenance to model benchmark performance.

Credibility is one of the quality dimensions in Zaveri et al.’s [15] survey of Linked Data quality assessment; PROV-CRED targets the representational gap rather than quality measurement. mUnc [16] takes a related but distinct approach: an RDFS vocabulary for representing arbitrary uncertainty theories generically (including their calculi, via LDScript), rather than committing to specific enrichment methods or benchmark-grounded reliability as PROV-CRED does. Where PROV-CRED ties credibility to the documented performance of the producing model, mUnc leaves the choice and semantics of the uncertainty theory itself open, to be fixed by whoever defines or selects the theory rather than derived from any producer’s track record. RDF-Star [7] (forthcoming as part of RDF 1.2) provides a syntactically clean mechanism for annotating individual triples without reification. It removes the storage objection to triple-level provenance but defines no semantics for what the annotations mean.

Content Credentials (C2PA) [17] attach a set of cryptographically signed manifests recording an asset’s capture and edit history to an image, video, or audio file. PROV-CRED and Content Credentials overlap in intent, recording how a piece of content came to exist so the history is verifiable rather than assumed. They diverge in grain and grounding: the C2PA specification explicitly disclaims value judgments, validating only that assertions are correctly formed and untampered rather than assessing a producing method’s reliability, whereas PROV-CRED annotates individual triples and grounds each triple’s credibility in the enrichment method’s documented benchmark performance.

2.3. Model Documentation

Model cards [18] and datasheets for datasets [19] standardize the documentation of model capabilities, limitations, and evaluation results as structured human-readable artefacts. Where a model card records benchmark performance as a static document, `provcred:Model` records it as a queryable graph node, linked to each triple the model produced via `provcred:ReliabilityAttribution`. `provcred:modelCard` on `provcred:Model` links the model node to the external model card URI, making the static document queryable as a graph property.

2.4. Computational Argumentation

Computational argumentation underlies applications that directly mediate public deliberation. Applications include claim detection, political debate analysis, fact checking, and legal argumentation [20]. Hi-ArG [21] organizes arguments as a two-level graph — intra-argument semantic relations (AMR-based) and inter-argument support/attack edges — and uses this structure to augment language model pre-training on argumentation tasks; the precision of each construction method is, however, unrepresented in the graph. Plenz et al. [22] enrich arguments with Contextualized Commonsense Knowledge Graphs (CCKGs), selecting ConceptNet paths by semantic similarity to connect premises to conclusions; edge weights encode argument relevance rather than source reliability, and the similarity model that selected each path is not recorded. PROV-CRED makes source reliability queryable where these approaches leave it absent or hand-assigned.

3. PROV-CRED

3.1. Design Principles

PROV-CRED is designed around three principles. First, each triple’s reliability score is derived from the producing model’s documented benchmark performance on the relevant task. For human-curated triples, credibility is anchored in cited sources or editorial review records. Second, enrichment method is the primary annotation axis: how a triple was produced determines the audit path when it is challenged. Third, reliability and credibility are distinct: reliability concerns empirical task performance; credibility concerns institutional authority over a producing model.

`provcred:ReliabilityAttribution` records benchmark-grounded technical performance, attached to each produced triple via PROV-O qualified attribution. `provcred:CredibilityAttribution` holds institutional accreditation, attached to the producing model. Both are subclasses of `prov:Attribution`; PROV-CRED extends PROV-O rather than replacing it.

3.2. Enrichment Method Classification

Four enrichment method classes are defined as subclasses of `provcred:KGenrichment` (itself a subclass of `prov:Activity`):

Table 1

Enrichment method classes and their audit properties.

Class	Reproducible?	Error traceable to
<code>HumanEnrichment</code>	Editorial-process dependent	Cited source, editorial review
<code>DeterministicEnrichment</code>	Yes	Explicit rule definition
<code>HeuristicEnrichment</code>	Yes, given same rules	Threshold or rule design
<code>ModelEnrichment</code>	Config-dependent	Weights, training data, prompt

Each class governs a distinct challenge path: `HumanEnrichment` (human-in-the-loop, HITL [23]) triples are challenged by contesting the cited source, the editorial judgment, or the community review process, `DeterministicEnrichment` by rebutting the rule definition, `HeuristicEnrichment` by contesting threshold or design choices, `ModelEnrichment` by contesting the model’s calibration on the relevant instance class. Reproducibility does not imply correctness: a deterministic rule can be systematically wrong on every firing, while a well-calibrated model-produced triple may be more credible in practice.

Reproducibility within `ModelEnrichment` depends on the `provcred:InferenceSetting`: KGE completion with fixed weights and a fixed `provcred:seed` is bit-for-bit reproducible, while LLM

inference with `temperature > 0` is not. This bit-for-bit guarantee concerns inference given already-fixed weights, not reproduction of those weights in the first place. Recent work shows that retraining an identical configuration, same hyperparameters and seed, on different GPU hardware produces models with prediction-level instability of the same magnitude as varying the seed itself, so the “fixed weights” this guarantee assumes are themselves a fragile reference point in practice [24]. A finer-grained subdivision into embedding and generative subclasses is left for future work.

3.3. Model Credentialing

Model credentialing is the process by which a model is rated on the task-specific benchmark most relevant to how it will be used in the enrichment pipeline. The rating is recorded once, at deployment time, and associated with the specific model version.

Representative task–benchmark pairings include:

- Named entity recognition: F_1 on a standard held-out evaluation set (e.g. CoNLL-2003 [25])
- Knowledge graph embedding: MRR and Hits@10 on FB15k-237 [26] (or a domain-matched benchmark, e.g. OGB-BioKG for biomedical KGs)
- Relation extraction: F_1 on SemEval-2010 Task 8 [27]
- LLM-based triple extraction: task-specific held-out evaluation as documented in the model’s technical report

A model carries a reliability *profile* by task type, not a single global score; the profile is the unit of credentialing, not the model as a whole.

Credentialing is valid only under *domain match*: the benchmark domain must approximate the deployment domain. A NER model evaluated on English news (CoNLL-2003) may not carry a valid credibility claim when applied to German cultural heritage records. The vocabulary records a `provcred:benchmarkDomain` property on `provcred:Model`; a mismatch between benchmark domain and deployment domain signals that benchmark-grounded reliability is inapplicable for that deployment context. Domain match is a precondition for the reliability score claim, not an open research problem: where it does not hold, the `provcred:ReliabilityAttribution` score is recorded but should not be treated as a valid warrant. `provcred:credentialedAt` records the date and time the evaluation was performed. It is placed on the attribution rather than the model because a single model may be re-evaluated on different benchmarks at different times.

3.4. Rated Triple Provenance

Each triple produced by a model records, via a PROV-O qualified attribution, the producing model, the task type, and the applicable reliability score. For most pipelines, one attribution record per enrichment activity (named-graph-level) is sufficient; triple-level RDF-Star annotation is appropriate for high-stakes deployments such as blood testing equipment KGs, where a misattributed sample-contraindication triple has direct patient safety consequences and per-triple audit trails are required.

The triple encodes a patient-safety constraint: coagulation analyzer v3 (`:coag-analyzer-v3`) must not be used with heparinized samples because heparin suppresses clotting-factor activity and interferes with the device’s coagulation assay, making it a claim where the reliability and credibility of the producing model are as consequential as the claim itself.

```
@prefix provcred: <https://prov-cred.org/ns/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix mls: <http://www.w3.org/ns/mls#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# Model credentialed on a biomedical KG benchmark, so benchmarkDomain
# approximates the deployment domain (domain-match precondition, 3.3).
:model-kge-transe-biokg-v2
```

```

a provcred:Model ;
provcred:modelCard <https://huggingface.co/...> ;
provcred:taskType      provcred:KGECompletion ;
provcred:trainingDataset provcred:OGBL-BioKG-train ;
provcred:evaluationDataset provcred:OGBL-BioKG-test ;
provcred:benchmarkDomain "biomedical KG" .

# Enrichment activity, typed by method: queryable as ModelEnrichment
# via prov:wasGeneratedBy (contribution 1).
:run-transe-2026-05-01
  a provcred:ModelEnrichment ;
  prov:wasAssociatedWith :model-kge-transe-biokg-v2 ;
  prov:endedAtTime "2026-05-01T09:14:00Z"^^xsd:dateTime ;
  provcred:inferenceSetting :inf-run-transe-2026-05-01 .

# Runtime parameters for this run (KGE: deterministic given weights).
:inf-run-transe-2026-05-01
  a provcred:InferenceSetting ;
  provcred:seed      42 ;
  provcred:scoreThreshold "0.80"^^xsd:decimal .

# Reliability attribution: the model's documented MRR on the
# task-relevant benchmark -- the standard KGE measure (3.3).
:attr-coag-analyzer-heparin
  a provcred:ReliabilityAttribution ;
  prov:agent      :model-kge-transe-biokg-v2 ;
  mls:specifiedBy provcred:MRR ;
  mls:hasValue    "0.7452"^^xsd:decimal ;
  provcred:basis  provcred:BenchmarkDerived ;
  provcred:credentialedAt "2026-03-15T00:00:00Z"^^xsd:dateTime .

# The rated triple.
<< :coag-analyzer-v3 :contraindicatedWith :heparinized-sample >>
  prov:wasGeneratedBy      :run-transe-2026-05-01 ;
  prov:wasAttributedTo     :model-kge-transe-biokg-v2 ;
  prov:qualifiedAttribution :attr-coag-analyzer-heparin .

```

Listing 1: RDF-Star annotation of a model-produced triple with PROV-CRED, recording the typed enrichment activity, the inference setting, and the benchmark-grounded reliability score.

The reliability score is the model's documented benchmark performance on the task-relevant measure (provcred:MRR for KGE completion; provcred:MicroF1 for named entity recognition). Evaluation measures are mls:EvaluationMeasure instances; the vocabulary includes MRR, Hits@1, Hits@10, Micro-F1, Macro-F1, and Precision. Scores are comparable only within the same measure; cross-measure ranking requires calibration that does not currently exist.

The provcred:trainingDataset property on provcred:Model distinguishes the data used to fit the model's weights from the held-out test split recorded in provcred:evaluationDataset.

Each produced triple links to a provcred:InferenceSetting node that records the runtime parameters under which the model ran. For KGE and other deterministic-given-weights models, the relevant parameters are provcred:seed (for reproducibility) and provcred:scoreThreshold (the minimum link-prediction score required to accept a candidate triple). For generative models, the config additionally carries provcred:temperature, provcred:topP, provcred:topK, and provcred:maxTokens. Attaching the config to the annotation rather than to the model instance reflects that inference parameters may vary per pipeline run, while model identity and credentialing remain fixed.

The annotation is *policy-level*: the score records expected accuracy across the method's task type. The warrant is “this method achieved an MRR of 0.7452 on the OGB-BioKG benchmark [28]”, not “this

triple is correct with probability 0.7452”. Benchmark performance at the method level is the minimum basis for a warranted per-triple reliability claim; the vocabulary makes that claim representable and queryable. A single SPARQL query retrieves the triple, its enrichment method, the applicable evaluation measure, and the benchmark basis.

The complete vocabulary — classes, properties, SKOS concept schemes for evaluation measures and dataset identifiers, and SHACL shapes — is available on GitHub². The vocabulary is minimal by design: it adds what PROV-O does not capture (method classification and benchmark-grounded reliability) without duplicating existing provenance semantics.

4. Motivating Application

HippoRAG [8] builds a knowledge graph as a long-term memory index for LLMs. An instruction-tuned LLM extracts open triples from a document corpus; retrieval encoders add synonym edges between noun phrases. Personalized PageRank (PPR) traverses the resulting KG for multi-hop retrieval. Medical diagnosis and legal case briefing are among the identified target use cases.

The KG is populated by two distinct `ModelEnrichment` activities: LLM-based open information extraction and encoder-based synonym linking, with no provenance distinction between them. A triple produced by the extraction LLM and a synonym edge from the retrieval encoder enter the same index; PPR traversal weights them equally. `provcred:ReliabilityAttribution` records the producing model and its benchmark score for each enrichment activity, making triple reliability queryable at traversal time. The annotation applies to the enrichment activity rather than the schema, so it covers the schemaless case. Where the hippocampal index feeds medical diagnosis, the question of which models are certified to populate it is what `provcred:CredibilityAttribution` is designed to make representable.

Metadata Aggregation. Metadata aggregators such as the Deutsche Digitale Bibliothek³ harvest records from hundreds of contributing institutions. DDB’s collection is a mixture of structured- and semi-structured data. Named Entity Recognition (NER) applied to title fields [29], using task-specific language models or LLMs, generates entity links to authority records: each recognized mention produces a `skos:exactMatch` or `skos:closeMatch` triple connecting the item to the German Authority File⁴ Werk URI. These triples are deposited alongside schema-mapped metadata in the same graph. Static quality measures assess completeness, field population, and link validity but record nothing about how the link was produced: a `skos:exactMatch` generated by a fine-tuned NER model and one produced by a deterministic string rule are indistinguishable at query time.

PROV-CRED annotates each link triple with the activity type that produced it (`DeterministicEnrichment` for rule-based matches, `HumanEnrichment` for institution-contributed records, `ModelEnrichment` for NER-derived links) and the producing model’s F_1 on the relevant NER benchmark via `provcred:ReliabilityAttribution`. A downstream application can distinguish authoritative rule-based matches from model-inferred ones, with the credibility basis recorded in the graph rather than documented only in pipeline configuration.

Argumentation Knowledge Graphs. Both Hi-ArG [21] and Plenz et al. [22] construct argumentation KGs through automated pipelines whose construction methods are not recorded alongside the resulting edges. `provcred:ReliabilityAttribution` makes the construction method and benchmark precision of each edge type queryable; downstream systems can consequently weight or filter edges by source reliability rather than treating all edges as uniform.

²<https://github.com/ISE-FIZKarlsruhe/provcred/blob/main/vocab/provcred.ttl>

³DDB, <https://www.deutsche-digitale-bibliothek.de/>

⁴Gemeinsame Normdatei (GND), https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

5. Research Agenda

One scope constraint, four open research problems, and one open provocation are identified below.

Calibration across methods (current scope constraint). Confidence scores produced by KGE models, LLMs, and heuristic extractors are not on a common scale. A KGE ranking score of 0.85 and an LLM self-reported confidence of 0.85 are not comparable without calibration. No standard cross-method calibration dataset for KG enrichment tasks is currently known to exist. This is not a deferred future problem but a current scope constraint: PROV-CRED’s immediate use cases are method-specific labeling, within-method ranking, and audit-path reasoning, all of which are available without cross-method calibration. Cross-method ranking of reliability scores is deferred; the vocabulary accommodates it when calibration mappings become available by linking `mls:EvaluationMeasure` instances to calibration records.

Self-evaluation bias. LLMs that produce and score their own triples exhibit correlated errors: the same features that produce an incorrect triple also produce an overconfident self-assessment. An external evaluation pipeline, relying on a separate model, a held-out dataset, or human review, is required to produce reliable scores for LLM-produced triples.

Annotation granularity (application-driven choice). Triple-level annotation via RDF-Star is precise; named-graph-level annotation, one record per enrichment activity, is practical and lower-overhead. Triple-level annotation’s storage cost is non-negligible at production scale: each annotated triple needs its own `prov:wasGeneratedBy`, `prov:wasAttributedTo`, and `prov:qualifiedAttribution` statements even when they all point to a single shared attribution record for the enrichment run. The appropriate granularity is determined by application requirements: triple-level annotation suits high-stakes use cases requiring per-triple audit trails; graph-level annotation suits non-critical enrichment pipelines. The vocabulary supports both; which granularity to apply in a given context is an open design question.

Retroactive provenance. Wikidata, DBpedia, and most large production KGs carry no enrichment method annotation. Bootstrapping enrichment method records from edit histories, revision logs, and system documentation is an open problem. Heuristic inference of enrichment method from edit metadata is plausible but has not been validated at scale. The Open World Assumption compounds this: an unannotated triple may represent human curation or simply absent provenance. `prov:Attribution` already provides the architectural slot for human contributions; formalizing attribution evidence for human-curated triples is left for future work.

Domain mismatch detection. Benchmark-grounded credibility is valid only when the benchmark domain approximates the deployment domain. A NER model evaluated on English news wire does not carry a valid credibility claim when applied to historical German records. No standard method exists to quantify or automatically flag domain mismatch; PROV-CRED records `provcred:benchmarkDomain` on the `provcred:Model` instance to make the mismatch detectable, but the threshold for inapplicability and any domain-adaptation correction remain open problems.

KG data accreditation governance (open provocation). `provcred:CredibilityAttribution` introduces an architectural slot for an institutional accreditation seal on a producing model, queryable via SPARQL alongside every triple the model generated. What the vocabulary does not answer is who has the authority to issue that seal, under what criteria, with what legal force, and how it is verified, renewed, or revoked. The W3C Verifiable Credentials Data Model [30] already secures exactly that kind of seal, an issuer’s claim about a subject, with a cryptographic proof and, optionally, a revocable status; its `relatedResource` property even lets a credential bind to external content via

a mandatory cryptographic digest. That digest is computed once, over content that must already exist at issuance time, precisely to detect later modification, not to cover content that does not yet exist. `provcred:CredibilityAttribution` targets that open-ended case instead: composing with `prov:wasGeneratedBy` so a model’s credential covers every future triple its enrichment activities generate, without rehashing anything each time. By analogy, the CE marking Declaration of Conformity framework existed before any specific notified body had authority to issue certificates. PROV-CRED proposes the representational framework; the accreditation infrastructure is the long-horizon research agenda.

6. Conclusion

PROV-CRED proposes a minimal vocabulary that addresses two gaps in heterogeneously enriched KGs: the reliability opacity problem, and the absence of representational infrastructure for institutional credibility attestation of KG enrichment models. Model credentialing grounds reliability scores in documented benchmark performance rather than self-reported confidence. Rated provenance inherits those scores for each triple produced, making the reliability basis visible at query time. The conference theme, *Bridging the Gap Between Curated and Induced Semantics*, presupposes that the gap is visible; for a system querying a modern KG, the difference between a schema-mapped triple and an LLM-predicted triple is not. Closing the gap remains a research problem; representing it is the prerequisite. The institutional accreditation question concerning who has authority to issue the seal that `provcred:CredibilityAttribution` holds, and under what criteria, remains open. The vocabulary makes the slot representable; the governance framework is the long-horizon provocation.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for drafting and editing under substantial human review and editorial oversight.

CRediT.

Mary Ann Tan: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Christian Nennemann:** Conceptualization, Validation, Writing – review & editing. **Genet Asefa Gesese:** Conceptualization, Validation, Writing – review & editing. **Harald Sack:** Project administration, Funding acquisition, Supervision, Writing – review & editing.

References

- [1] T. Petrova, B. Bliznioukov, A. Puzikov, R. State, From multi-agent systems and the semantic web to agentic AI: A unified narrative of the web of agents, 2026. URL: <https://arxiv.org/abs/2507.10644>. `arXiv:2507.10644`.
- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, 2025. URL: <https://arxiv.org/abs/2404.16130>. `arXiv:2404.16130`.

- [4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* 55 (2023) 1–38. URL: <http://dx.doi.org/10.1145/3571730>. doi:10.1145/3571730.
- [5] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering* 36 (2024) 3580–3599. URL: <http://dx.doi.org/10.1109/TKDE.2024.3352100>. doi:10.1109/tkde.2024.3352100.
- [6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. J. Hruschka, T. M. Mitchell, Toward an architecture for never-ending language learning, in: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI'10*, AAAI Press, 2010, pp. 1306–1313.
- [7] O. Hartig, P.-A. Champin, G. Kellogg, A. Seaborne, RDF 1.2 concepts and abstract syntax, W3C Working Draft, 2024. URL: <https://www.w3.org/TR/rdf12-concepts/>.
- [8] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, Y. Su, Hipporag: Neurobiologically inspired long-term memory for large language models, 2025. URL: <https://arxiv.org/abs/2405.14831>. arXiv:2405.14831.
- [9] T. Lebo, S. Sahoo, D. McGuinness, et al., PROV-O: The PROV ontology, W3C Recommendation, 2013. URL: <https://www.w3.org/TR/prov-o/>.
- [10] W. Wu, H. Li, H. Wang, K. Q. Zhu, Probase: a probabilistic taxonomy for text understanding, in: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12*, Association for Computing Machinery, New York, NY, USA, 2012, p. 481–492. URL: <https://doi.org/10.1145/2213836.2213891>. doi:10.1145/2213836.2213891.
- [11] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, G. Weikum, Yago2: exploring and querying world knowledge in time, space, context, and many languages, in: *Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11*, Association for Computing Machinery, New York, NY, USA, 2011, p. 229–232. URL: <https://doi.org/10.1145/1963192.1963296>. doi:10.1145/1963192.1963296.
- [12] J. Cao, J. Fang, Z. Meng, S. Liang, Knowledge graph embedding: A survey from the perspective of representation spaces, *ACM Comput. Surv.* 56 (2024). URL: <https://doi.org/10.1145/3643806>. doi:10.1145/3643806.
- [13] T. Wu, K. Yao, W. Li, G. Qi, Y. Yu, N. Zhao, R. Zhang, P. Duan, Triple confidence measurement in knowledge graph with multiple heterogeneous evidences, *World Wide Web* 27 (2024) 70.
- [14] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, *Information Services and Use* 30 (2010) 51–56. URL: <https://doi.org/10.3233/ISU-2010-0613>. doi:10.3233/ISU-2010-0613. arXiv:<https://doi.org/10.3233/ISU-2010-0613>.
- [15] P. Hitzler, A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, *Semantic Web* 7 (2015) 63–93. URL: <https://doi.org/10.3233/SW-150175>. doi:10.3233/SW-150175.
- [16] A. E. A. Djebri, A. G. Tettamanzi, F. Gandon, Linking and negotiating uncertainty theories over linked data, in: *Companion Proceedings of The 2019 World Wide Web Conference, LDOW/LDDL Workshop*, 2019, pp. 859–865. doi:10.1145/3308560.3317074.
- [17] Coalition for Content Provenance and Authenticity, Content credentials: C2PA technical specification, Technical Specification, v2.4, 2026. URL: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html.
- [18] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, p. 220–229. URL: <http://dx.doi.org/10.1145/3287560.3287596>. doi:10.1145/3287560.3287596.
- [19] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets, 2021. URL: <https://arxiv.org/abs/1803.09010>. arXiv:1803.09010.
- [20] J. Lawrence, C. Reed, Argument mining: A survey, *Computational Linguistics* 45 (2019) 765–818. URL: <https://aclanthology.org/J19-4006/>. doi:10.1162/coli_a_00364.
- [21] J. Liang, R. Ye, M. Han, Q. Zhang, R. Lai, X. Zhang, Z. Cao, X. Huang, Z. Wei, Hi-ArG: Exploring the integration of hierarchical argumentation graphs in language pretraining, in: H. Bouamor, J. Pino,

- K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, pp. 14606–14620. URL: <https://aclanthology.org/2023.emnlp-main.902/>. doi:10.18653/v1/2023.emnlp-main.902.
- [22] M. Plenz, J. Opitz, P. Heinisch, P. Cimiano, A. Frank, Similarity-weighted construction of contextualized commonsense knowledge graphs for knowledge-intense argumentation tasks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6130–6158. URL: <https://aclanthology.org/2023.acl-long.338/>. doi:10.18653/v1/2023.acl-long.338.
- [23] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, Power to the people: The role of humans in interactive machine learning, *AI Mag.* 35 (2014) 105–120. URL: <https://doi.org/10.1609/aimag.v35i4.2513>. doi:10.1609/aimag.v35i4.2513.
- [24] G. Méroué, F. Gandon, P. Monnin, Link Prediction or Perdition: The Seeds of Instability in Knowledge Graph Embeddings, Springer-Verlag, Berlin, Heidelberg, 2026, p. 198–216. URL: https://doi.org/10.1007/978-3-032-25156-5_11. doi:10.1007/978-3-032-25156-5_11.
- [25] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. URL: <https://aclanthology.org/W03-0419/>.
- [26] K. Toutanova, D. Chen, Observed versus latent features for knowledge base and text inference, in: *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, Association for Computational Linguistics, Beijing, China, 2015, pp. 57–66. URL: <https://aclanthology.org/W15-4007/>. doi:10.18653/v1/W15-4007.
- [27] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, in: *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 33–38. URL: <https://aclanthology.org/S10-1006/>.
- [28] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, J. Leskovec, Open graph benchmark: Datasets for machine learning on graphs, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 22118–22133. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/fb60d411a5c5b72b2e7d3527cfc84fd0-Paper.pdf.
- [29] M. A. Tan, S. Jiang, H. Sack, How to Turn Card Catalogs into LLM Fodder, in: *Workshop on Deep Learning and Linguistic Linked Data*, 2024.
- [30] M. Sporny, D. Longley, D. Chadwick, I. Herman, Verifiable credentials data model v2.0, W3C Recommendation, 2025. URL: <https://www.w3.org/TR/2025/REC-vc-data-model-2.0-20250515/>.