

AI-DRIVEN METADATA MANAGEMENT: A GOVERNANCE ARCHITECTURE FOR PROBABILISTIC SEMANTIC ENRICHMENT

Computer Science

Aditya Ranjan

FNU Independent Researcher

ABSTRACT

Modern data platforms span warehouses, lakehouses, streaming pipelines, and retrieval-augmented applications, creating pressure to enrich metadata automatically without weakening governance. This study examines how probabilistic semantic enrichment can coexist with versioned, attributable, and auditable metadata. A platform-neutral control architecture is proposed that separates observed metadata from AI-generated proposals and makes promotion conditional on provenance, confidence, deterministic evidence, version context, and human review. The design is evaluated with public New York City Taxi and Limousine Commission (NYC TLC) schema metadata in two controlled experiments. In a leakage-safe semantic-labeling test that holds out complete field identities, a TF-IDF and logistic-regression baseline reaches 0.667 accuracy and 0.535 macro-F1 on 72 held-out records across five classes. Location and temporal fields generalize well, whereas unseen identifier and operational fields are often confused with financial metadata, showing the limits of text-only authority. In a TLC-seeded synthetic catalog, sparse semantic retrieval from 10,000 to 100,000 entities raises median latency from 1.75 ms to 18.29 ms in the documented local environment. A documented 2025 schema change illustrates lineage-aware impact analysis. The findings support a pattern in which AI expands metadata coverage while deterministic governance retains authority over trusted metadata.

KEYWORDS

Metadata Management; Data Governance; Semantic Enrichment; Data Lineage

1. INTRODUCTION

A data catalog was once judged by whether engineers could find a table and understand its columns. Today the same asset may feed SQL workloads, streaming jobs, dashboards, machine-learning features, vector retrieval, and AI agents. A wrong description is an inconvenience; a wrong sensitivity tag, lineage edge, or business meaning can change access decisions and downstream behavior.

This paper treats metadata as a governed control plane rather than a documentation layer. Automated enrichment increases semantic coverage, yet authoritative metadata requires stable identity, version context, provenance, and accountable control. Treating model output as ordinary catalog truth improves coverage but weakens auditability. The study therefore asks how probabilistic enrichment can be incorporated without collapsing the distinction between inference and authority.

Four contributions follow: (a) a platform-neutral control architecture that separates observed metadata from enrichment proposals and governs promotion; (b) a leakage-safe evaluation in which complete field identities are held out; (c) evidence requirements derived from class-specific failures; and (d) a retrieval-scaling experiment and schema-evolution case linking semantic candidate generation to lineage-aware impact analysis.

2. Related Work

ISO/IEC 11179-1:2023 defines a framework for metadata registries and treats metadata as descriptions of data (ISO/IEC, 2023). The W3C DCAT 3 Recommendation provides an RDF vocabulary for interoperable catalogs, with support for versioning and dataset series (W3C, 2024). OpenLineage standardizes runtime and design-time lineage events across jobs and datasets (OpenLineage, 2026), and Eichler et al. (2021) proposed a generic metadata model for data lakes.

Open-source catalogs provide mature foundations. Apache Atlas supports classifications, propagation through lineage, search, and lineage APIs (Apache Software Foundation, 2026), and the OpenMetadata sizing runbook tests scale from 10K to 5M entities while tracking throughput, p95 latency, and error rate (OpenMetadata, 2026). Metadata-driven design has also been applied to ingestion: the MIND pattern uses metadata to adapt ingestion behavior, schema updates, and source integration in cloud big-data platforms (Rucco et al., 2026).

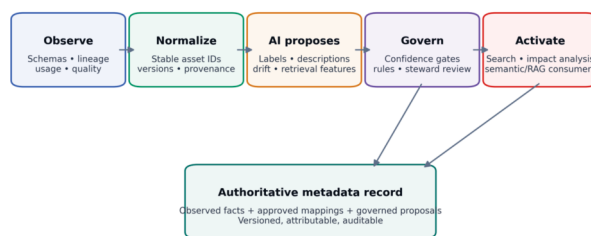
Storage, discovery, and lineage are therefore established problems. The unresolved question is what evidence, and governance boundary should precede promotion of probabilistically inferred metadata into trusted metadata, because deterministic governance does not scale semantic authoring and model confidence alone cannot establish authority.

3. Proposed Architecture

The governing principle is conservative: AI may propose, rank,

summarize, or flag, but authoritative metadata remains attributable to observed source facts, deterministic rules, approved mappings, or reviewed decisions. This matches the risk-management orientation of the NIST AI RMF 1.0 (Tabassi, 2023). Figure 1 shows the resulting control loop.

Governed AI-driven metadata control loop



Design rule: AI may propose or rank; promotion into trusted metadata is controlled by evidence and governance.

Figure 1: Governed AI-driven Metadata Control Loop

Connectors first capture schemas, owners, usage, quality signals, and lineage events; normalization then assigns stable asset identities and versions. Only afterwards does the enrichment service operate. Each proposal records the proposed value, confidence, model or rule version, source features, timestamp, and asset version. Governance gates apply risk-dependent policy: search synonyms can be activated with low friction, descriptions may require a confidence threshold, and sensitivity, retention, or access-impacting metadata requires deterministic corroboration or steward approval. Search, impact analysis, semantic layers, RAG systems, and agents consume only governed metadata, so a similarity score never silently becomes a security or compliance decision.

4. Data and Methods

4.1 Data

The case study uses NYC TLC Trip Record Data, published monthly with a typical two-month delay (New York City Taxi and Limousine Commission [NYC TLC], 2026). The TLC page states that from 2025 onward a `cbd_congestion_fee` column was added to the Yellow, Green, and High Volume FHV datasets to reflect congestion-pricing charges. This documented change serves as the schema-drift event. No passenger-level findings are claimed.

4.2 Semantic-enrichment Experiment

Twenty TLC-derived field names and descriptions were assigned to five coarse classes: financial, identifier, location, operational, and temporal. Twelve controlled textual variants per field gave 240 records. A TF-IDF (unigram and bigram) and class-balanced logistic-regression classifier was trained. To prevent paraphrase leakage, six complete field identities (`fare_amount`, `tolls_amount`, `RatecodeID`, `PULocationID`, `trip_distance`, and `tppep_dropoff_datetime`) were held out, leaving 168 training and 72 test records, with every class present in

the test set. Accuracy and macro-F1 are reported.

4.3 Retrieval Benchmark

TLC vocabulary seeded synthetic catalogs of 10,000, 50,000, and 100,000 entities, enlarged with distinct asset identifiers and partition markers. A TF-IDF index was built for each size, and five queries were run 20 times each after warm-up, computing sparse similarity and selecting the top five candidates. The benchmark excludes network calls, graph traversal, authentication, and concurrent traffic, so it is a component-level baseline, not a test of any catalog product.

4.4 Environment

Experiments ran on Linux 6.18.35 (x86_64) with Python 3.13.5, NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, and random seed 42. CPU model, memory, and container limits were not exposed, so latencies are observations from this environment rather than hardware-normalized claims.

5. RESULTS

5.1 Semantic Classification

The held-out test achieved 0.667 accuracy and 0.535 macro-F1 across 72 records (Table 1). Location and temporal fields reached an F1 of 1.000, and financial fields 0.676. Held-out identifier and operational fields were not recovered; most were absorbed into the financial class, lowering its precision to 0.511.

Table 1: Per-class Results on the 72 Held-out Records

Class	Precision	Recall	F1	Support
Financial	0.511	1.000	0.676	24
Identifier	0.000	0.000	0.000	12
Location	1.000	1.000	1.000	12
Operational	0.000	0.000	0.000	12
Temporal	1.000	1.000	1.000	12

Text-only models recognize classes with distinctive vocabulary, such as locations and timestamps, but coarse classes become ambiguous when an unseen field lacks context. Production enrichment should therefore treat text as one signal among several, adding datatype, neighboring fields, lineage, ownership, statistics, and steward review.

5.2 Schema Drift

A schema snapshot creates a version fingerprint, and set comparison identifies the added `cbd_congestion_fee` field. The event is attached to the dataset version, the field is queued for classification with provenance, and lineage traversal flags consumers such as fare-reconciliation logic, dashboards, features, and RAG indexes for review. Recording the 2025 validity context protects historical comparability.

5.3 Retrieval Performance

Table 2 shows median latency rising from 1.75 ms at 10,000 entities to 18.29 ms at 100,000, with a p99 of 20.17 ms. Index build time rose from 0.142 s to 1.033 s. These values isolate retrieval mechanics and are not end-to-end catalog performance.

Table 2: Retrieval Latency and Index-build Time by Catalog Size

Entities	Build (s)	p50 (ms)	p95 (ms)	p99 (ms)
10,000	0.142	1.75	1.85	2.02
50,000	0.432	8.79	9.37	10.21
100,000	1.033	18.29	19.64	20.17

6. DISCUSSION AND LIMITATIONS

A weaker but leakage-safe result is more informative than a perfect score produced by near-duplicate training and test text. The 0.535 macro-F1 shows that enrichment cannot rely on name and description alone, and confidence should be calibrated by class and evidence type: a location label supported by a taxi-zone foreign key differs qualitatively from a financial label inferred only from wording.

For AI applications, metadata also constrains machine retrieval. A RAG system that retrieves a similar table without checking certification, freshness, policy, or version can answer from the wrong asset. A safer pattern is two-stage retrieval: semantic candidate generation followed by deterministic metadata filters and lineage-aware ranking. Adoption should place AI after identity, versioning, and lineage, moving from deterministic controls to AI shadow mode, then controlled activation of reversible low-risk outputs.

Limitations are material. The corpus is small, domain-specific, and

coarse, and its text variants are controlled rather than independently annotated production metadata. External validity is limited to one public data domain. The catalog-scale workload is synthetic, with only its vocabulary grounded in TLC metadata, and latency depends on the environment. The study supports the proposed control boundary and documents failure modes; it does not establish universal accuracy, production latency, or superiority over commercial catalogs.

7. CONCLUSION

Probabilistic enrichment can expand semantic metadata coverage, but model confidence alone is not a sufficient basis for authoritative metadata. Under identity-level holdout, the text-only baseline generalized for location and temporal metadata but failed on unseen identifier and operational fields, motivating the separation of AI proposals from governed authority. The TLC schema change shows how versioned metadata turns a new field into an impact-analysis event. Provenance, confidence, version context, lineage, deterministic checks, and human review should decide when proposals become trusted metadata. Future work should use several public schemas, multiple annotators, and cross-domain evaluation.

DECLARATIONS

The study uses public NYC TLC metadata; the controlled corpus, scripts, environment specifications, and raw outputs are available from the author on request. The author received no specific funding and declared no competing interests. AI tools supported language refinement, document organization, source discovery, fact-checking, and initial figure-layout drafting; the author reviewed the content, verified results against the reproducibility materials, and takes full responsibility for it.

REFERENCES

1. Apache Software Foundation. (2026). *Apache Atlas documentation: Classifications, lineage, search and REST APIs*. <https://atlas.apache.org/>
2. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2021). Modeling metadata in data lakes—A generic model. *Data & Knowledge Engineering*, 136, Article 101931. <https://doi.org/10.1016/j.datak.2021.101931>
3. ISO/IEC. (2023). *Information technology—Metadata registries (MDR)—Part 1: Framework* (ISO/IEC 11179-1:2023, 4th ed.). International Organization for Standardization.
4. New York City Taxi and Limousine Commission. (2026). *TLC trip record data*. <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
5. OpenLineage. (2026). *Object model and specification*. <https://openlineage.io/docs/spec/object-model/>
6. OpenMetadata. (2026). *Cluster sizing runbook* [Documentation]. GitHub, `open-metadata/OpenMetadata, bin/distributed-test/CLUSTER-SIZING-RUNBOOK.md`.
7. Rucco, C., Longo, A., & Saad, M. (2026). MIND: A metadata-driven INgestion design pattern for efficient data ingestion. *Big Data Research*, 43, Article 100574. <https://doi.org/10.1016/j.bdr.2025.100574>
8. Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
9. W3C. (2024). *Data catalog vocabulary (DCAT)—Version 3* (W3C Recommendation). <https://www.w3.org/TR/vocab-dcat-3/>