Identity, Equivalence, and Reconciliation across Heterogeneous Data Systems

Matthew Long
The YonedaAI Collaboration
YonedaAI Research Collective
Chicago, IL
matthew@yonedaai.com · https://yonedaai.com

23 July 2026

Abstract

Two records can look alike without naming the same thing. They can also name the same thing while looking quite different. In an integration platform, a mistaken match can quietly change joins, counts, materialized views, and writes. This paper specifies a provisional identity and reconciliation layer for CatDB. The design keeps four artifacts separate: immutable source-local references, observed evidence, versioned method results, and authority-bearing reconciliation decisions. A canonical identity is a versioned projection of admitted decisions under a named identity profile. It is not an ontological fact inferred from a score.

The model records source-local identity, canonical anchors, candidate links, possible and probable equivalence, confirmed sameness, confirmed difference, contradiction, supersession, retraction, and versioned identity as distinct states and artifacts. Only effective `confirmed-same` decisions may enter an equivalence projection. Vector similarity and approximate nearest-neighbor search may generate candidates or features, but they cannot establish semantic identity by themselves. The paper also specifies scoped authority and source ownership, along with bitemporal replay, provenance, logical reversal, downstream invalidation, and strict versus partial federation behavior.

The contribution remains an ENGINEERING HYPOTHESIS. Probabilistic record linkage, merge/purge, blocking, collective resolution, declarative matching, human-in-the-loop matching, identity uncertainty, Semantic Web equality, provenance, and large-scale entity-matching systems are established prior work. A repository audit found no accepted identity profile, `catdb-identity` crate, identity fixture family, Haskell identity oracle, Lean identity theorem, reconciliation state machine, calibrated matcher, or P7 experiment. Every CatDB runtime claim is therefore OPEN or OBSTRUCTED. The strongest plausible systems contribution is narrower: an auditable, typed, versioned decision layer that preserves uncertainty and evidence, applies scoped authority, supports admitted reversals, and exposes their consequences across mappings, federation, materialization, provenance, and workspaces. The implementation and experimental gates defined here must determine whether that combination is useful.

Contents

1 Introduction

1.1 One customer, three records, and four different questions

Consider a retailer that runs a sales database, a billing service, and a support system. The sales database has customer `crm:4821`; billing has account `billing:773`; support has contact `ticketing:51`. The names are similar, and two records share an email address. The third record was created after that address changed, but its delivery address and purchase history are compatible with the first record.

An integration engineer might ask whether these records belong to the same household, natural person, billing account, or legal contracting party. Those are different relations, not four phrasings of one relation. The first may group family members; the third may deliberately separate accounts held by one person; the fourth may depend on jurisdiction and effective date. Even the apparently simple question "same person" depends on source-key reuse, time, evidence quality, and authority.

Suppose a vector index places the three records near each other. The result is useful because the system has avoided comparing every record with every other record. It is not a decision. The embedding was trained for a declared objective, uses a particular projection of the source fields, and may return semantically similar but distinct people. Approximate search may also omit a true neighbor. A classifier score is another method result. Human review adds a judgment, and a source owner may supply an authoritative crosswalk. The final merge must not overwrite any of them.

The risk extends beyond a wrong label. If the system treats a likely match as equality, a federated query may undercount customers. A materialized view may combine balances, a workspace merge may propagate the decision, and a write-through policy may update the wrong source. Removing one pair edge later may neither reconstruct the old partition nor reveal where synthesized values belong.

This paper examines the semantic boundary before those actions. It asks:

> How can identity evidence remain explicit without collapsing uncertainty?

1.2 The proposed answer

Within an immutable identity profile, CatDB should represent separately:

1. source-local references and their key/snapshot regimes;
2. positive, negative, missing, and dependent evidence;
3. deterministic outcomes, likelihood ratios, probabilities, similarities, ranks, and human judgments;
4. scoped authority and source ownership;
5. proposals, approvals, rejections, contradictions, supersessions, and reversals; and
6. the canonical partition projected from effective confirmed decisions at a declared valid-time and system-time cut.

The profile fixes the entity type and exact identity question, as well as the context, time policy, admitted evidence and methods, decision states, authority, constraints, cluster policy, disclosure, retention, downstream obligations, and normalization. A canonical anchor names one versioned decision history. It neither replaces its member source references nor claims universal, metaphysical, or legal identity.

1.3 Research contribution and evidence boundary

Record linkage has a mature statistical foundation [8, 13]. Duplicate detection, match/merge systems, identity uncertainty, collective resolution, declarative constraints, crowdsourced review, and learned matching all precede this proposal [2–6, 10, 12, 15, 16, 19, 20, 23]. Approximate vector systems establish efficient similarity retrieval, not identity [14, 17]. OWL gives `owl:sameAs` strong equality semantics, while SKOS offers weaker concept-mapping relations [11, 25, 26]. W3C PROV supplies an interchange vocabulary for derivation, attribution, delegation, revision, and invalidation, but not this paper's decision state machine [27].

The proposed delta is not a new matcher. It is the composition of identity artifacts with CatDB's typed mappings, provenance, versioning, federation, and materialization boundaries. That composition remains unimplemented. This manuscript provides a candidate contract, examples, falsifiers, a prior-art boundary, an implementation roadmap, and a preregistered evaluation design. It reports no CatDB identity accuracy, calibration, scalability, reversibility, or machine-checked theorem.

1.4 Paper organization

Section 2 fixes the evidence vocabulary and current audit. Section 3 gives the hypothesis and falsification boundary. Section 4 defines the identity profile and artifact types. Section 5 defines uncertain and effective states. Section 6 separates deterministic, probabilistic, vector, and collective methods. Sections 7 to 9 treat authority, provenance, reversal, and integration. Section 10 states the formal targets without presenting them as proved. Section 11 follows one reconciliation case. Section 12 locates the proposal in prior work. Sections 13 and 14 define the build and experiment program. The final sections retain limitations, open questions, and exact nonclaims.

2 Evidence vocabulary and repository audit

2.1 Binding labels

The labels in Table 1 apply to every result-like statement. They prevent a design sketch from being narrated as a system result and a program run from being narrated as proof of real-world identity.

Table 1: Evidence labels used in this paper.

Label	Meaning
SUPPORTED BY PRIOR LITERATURE	A primary paper, standard, or official system artifact establishes the claim within its assumptions.
ENGINEERING HYPOTHESIS	A proposed CatDB behavior that requires implementation and experiment.
OPEN	Definitions, code, fixtures, experiments, or review are not complete.
OBSTRUCTED	A prerequisite semantic boundary is missing, so implementation or testing would be premature.
OPEN CONJECTURE	A mathematical statement remains unproved.
SPECULATIVE	The claim is too underspecified for a stronger label.

No P7 claim is labelled `proved`, `machine-checked`, `computational`, or `experimentally supported`. Those labels are reserved for exact future artifacts. Existing Rust, Haskell, and Lean work on categorical identity paths or identity schema mappings concerns algebraic identity operations, not entity identity.

2.2 What exists

The repository contains a P7 manifest and dependency edges, a proposed shared claim, terminology, a primary-source map, a detailed evidence dossier, and a smaller accepted semantic core for schema paths and mappings. The dossier also defines candidate identity fixture families, implementation phases, formal targets, benchmark questions, and review gates.

These are author-side research artifacts. They establish that the proposal has a falsifiable scope. They do not establish any identity decision, canonical partition, reversal, or performance property.

2.3 What does not exist

At the evidence cut used for this paper, the repository has no:

- accepted identity profile or profile validator;
- identity fixture family or normalized result manifest;
- `catdb-identity` crate;
- deterministic entity-resolution rule engine;
- Fellegi–Sunter evaluator or calibrated learned matcher;
- vector-candidate adapter with decision-path separation;
- authority or source-ownership policy implementation;
- human-review artifact model or reconciliation state machine;
- canonical projection, bitemporal replay, or reversible merge;
- downstream invalidation implementation;
- Haskell identity oracle or Lean identity theorem; or
- P7 benchmark, security analysis, or human study.

The overall P7 status is therefore OBSTRUCTED. Some isolated tasks are OPEN, while cross-layer tasks depend on unfinished mapping, materialization, workspace, policy, or provenance semantics.

2.4 Dependency boundary

P7 depends on three upstream paper programs:

1. P1 supplies immutable typed entities, values, paths, equations, and stable semantic references.
2. P4 supplies the record-preserving interpretation of source-to-domain mappings. A field mapping must not become an identity assertion.

3. P13 supplies provenance identities for sources, observations, methods, policies, actors, decisions, revisions, and derived products.

P5 is an execution dependency when federated answers reconcile records. P6 is needed to invalidate materializations after reversal. P7 in turn blocks safe write propagation in P8 and semantic workspace merge in P9.

3 Research question, hypothesis, and falsification

3.1 Primary hypothesis

Let R be immutable source-record references, P a versioned identity profile, E typed evidence observations, M versioned deterministic or probabilistic methods, A scoped authority policies, D reconciliation decisions, and C canonical anchors. For entity type τ, context χ, and validity interval ν, the proposed operation is

$$\mathrm{reconcile}_P : (R, E, M, A, D) \longrightarrow (\widehat{D}, C, \mathcal{O}),$$

where $\widehat{D}$ contains proposed or revised decisions and $\mathcal{O}$ contains validation, review, provenance, disclosure, invalidation, and execution obligations.

The hypothesis is that a finite, explicitly delimited profile can preserve source references, keep evidence and decisions distinct, prevent uncertain edges from entering equality, apply scoped authority, retain contradictions, support an admitted reversible fragment, reproduce historical projections, and expose downstream consequences.
Evidence status. ENGINEERING HYPOTHESIS.

3.2 Strongest defensible first target

The first useful result should be deliberately small:

- two entity types, for example `Person` and `Organization`;
- four sources with immutable record and snapshot references;
- exact and composite-key deterministic rules;
- one Fellegi–Sunter-style probabilistic matcher;
- one supervised pair classifier;
- one exact and one approximate vector candidate generator;
- explicit candidate, possible, probable, confirmed-same, confirmed-different, contradicted, superseded, and retracted states;
- one source-owner and one cross-domain approval policy;
- append-only bitemporal decision events;
- one admitted merge, split, and supersession fragment;

- P13-compatible provenance projection;
- deterministic Rust/Haskell agreement on shared fixtures; and
- selected Lean invariants over transitions and projections.

This target excludes universal identity, unrestricted biometric linkage, legal identity across jurisdictions, automatic cross-tenant write-back, and claims that one model transports across domains.

3.3 Falsification conditions

The hypothesis must be narrowed or rejected if any supported profile allows:

1. destruction of a source record or source-local identifier during canonicalization;
2. candidate generation to alter the effective canonical partition;
3. possible or probable edges to be transitively closed as equality;
4. confirmed-same across types or scopes without an explicit coercion;
5. a cannot-link pair to share an accepted anchor;
6. a decision without exact evidence, method, feature, threshold, policy, authority, and source-snapshot revisions;
7. an uncalibrated score to be displayed as a probability;
8. nondeterministic normalized output for the same immutable input;
9. nonreproducible replay at the same system-time cut;
10. a claimed supported reversal that cannot reconstruct the prior partition;
11. a reversed decision to leave an affected artifact presented as current;
12. an actor to approve outside recorded authority;
13. source ownership to disappear after canonicalization;
14. approximate-search misses to be absent from candidate-recall reporting;
15. missing evidence to be treated as negative evidence;
16. partial source availability to be presented as complete;
17. model drift to leave old outputs current without revalidation; or
18. claimed workflow benefit to vanish after review, provenance, calibration, candidate, repair, and reversal costs are counted.

These falsifiers are more important than the product vocabulary. A system that stores the proposed record shapes but violates them would not satisfy the P7 hypothesis.

4 A profile-relative identity model

4.1 Source-local identity

Definition 4.1 (Source-local reference)**.** A source-local reference is a tuple

$$\mathsf{SourceRef} = (\text{source}, \text{namespace}, \text{record key}, \text{snapshot}).$$

Its interpretation is fixed by a source regime that declares key stability, reuse, deletion, tombstone, alias, and snapshot behavior.

A source-local reference makes an observed record addressable. It does not make the key global, unrecycled, correct, or authoritative. Nor does it say that equal keys in two namespaces denote the same entity. A durable source reference is therefore a prerequisite for reconciliation, not its outcome.

4.2 Canonical identity anchor

Definition 4.2 (Canonical anchor)**.** A canonical anchor is an immutable identifier

$$\mathsf{CanonicalRef} = (\text{identity profile}, \text{anchor ID}, \text{revision})$$

for a versioned canonicalization history.

The current membership of an anchor is derived from effective decisions at a declared time cut. Member source references remain intact. The anchor is not a natural key, an `owl:sameAs` assertion by default, a guarantee that every member is correct, or evidence that no unobserved duplicate exists.

4.3 Identity profile

Definition 4.3 (Identity profile)**.** An identity profile is an immutable, content-addressed contract

$$P = (\tau, q, \chi, V, S, E, M, K, D, A, H, G, X, L, N),$$

where τ is the entity type, q the exact identity question, χ the context, V the valid-time policy, S source regimes, E evidence types, M methods, K candidate policy, D decision states, A authority policy, H hard constraints, G cluster policy, X disclosure and retention policy, L downstream obligations, and N normalization.

The identity question is first-class. "Same household," "same person," "same patient chart," "same customer account," and "same legal organization" denote different relations. Decisions under different profiles are not automatically comparable or composable.
Evidence status. The identity-profile contract is OBSTRUCTED pending P4 and P13 reference semantics..

4.4 Four artifact kinds

P7 requires four typed artifacts rather than one overloaded match edge.

Definition 4.4 (Evidence observation)**.** An evidence observation records subjects, source artifact and snapshot, observed value or digest, evidence kind, polarity, independence group, sensitivity, valid interval, and provenance. It records what was observed, not what decision should be effective.

Definition 4.5 (Method result)**.** A method result records the method revision, input evidence, candidate pair or cluster, output kind and value, calibration population, threshold, uncertainty, feature digest, generation time, and provenance.

An output kind may be a deterministic rule outcome, likelihood ratio, estimated posterior, classifier score, calibrated probability, vector distance, candidate rank, cluster proposal, or `unsupported`. These forms are not interchangeable.

Definition 4.6 (Reconciliation decision)**.** A reconciliation decision records a profile, subjects, prior decisions, proposed and effective states, evidence and method results, policy revision, actor, authority scope, approvals, reason, valid interval, system time, supersession, reversal, obligations, and provenance.

Definition 4.7 (Canonical projection)**.** For profile P, valid-time cut t_v, and system-time cut t_s,

$$\mathsf{Project}(P, t_v, t_s)$$

is the normalized partition and conflict report computed from decisions visible at t_s and effective at t_v.

The proposed projection excludes retracted or superseded decisions according to profile rules, enforces confirmed-different constraints, retains unresolved references separately, reports contradictions, and is independent of incidental input enumeration.

4.5 Evidence is not one scalar

The model keeps apart:

- strength of an observation for or against sameness;
- a method's reported confidence;
- calibration of probabilities on a declared population;
- measured source reliability;
- institutional source authority;
- ownership of the source record;
- reviewer authority;
- operational decision status; and

- truth in the modeled world.

Truth-discovery work can model source reliability and dependence [7, 28]. Proper scoring rules can evaluate forecasts [9]. Neither statistical result grants an institution authority over a record or turns a probability into an operational decision.

5 Uncertainty, equivalence, and contradiction

5.1 Required states

Table 2 is the minimum state vocabulary. It deliberately refuses to make uncertainty transitive.

Table 2: Decision states and their effect on canonical identity.

State	Meaning	Canonical effect
`candidate`	Retrieved for detailed comparison.	none
`possible`	Evidence makes sameness plausible but insufficient.	none
`probable`	A calibrated method crosses a profile threshold.	none by default
`confirmed-same`	Authorized sameness decision within profile and time.	equivalence edge
`confirmed-different`	Authorized cannot-link decision.	merge constraint
`contradicted`	Effective evidence or decisions conflict.	blocks or routes review
`superseded`	A later decision replaces operational effect.	determined by successor
`retracted`	Decision withdrawn while history remains.	none
`expired`	Time or policy ended operational effect.	none

The three Fellegi–Sunter action regions motivate match, possible-match, and nonmatch handling [8]. They do not establish the full state machine above, and P7 must not attribute it to that theory.

5.2 Confirmed equivalence

Within one profile and time cut, only the effective `confirmed-same` relation is intended to generate an equivalence projection:

$$r \sim r,$$
$$r_1 \sim r_2 \Longrightarrow r_2 \sim r_1,$$
$$r_1 \sim r_2 \wedge r_2 \sim r_3 \Longrightarrow r_1 \sim r_3.$$

Candidate, possible, probable, vector-neighbor, and shared-evidence relations need not obey these laws. Their graph may be asymmetric, nontransitive, incomplete, or inconsistent.

Requirement 5.1 (No uncertain closure)**.** No `candidate`, `possible`, or `probable` edge may change a canonical partition unless a separate authorized transition creates an effective `confirmed-same` decision.

Evidence status. OBSTRUCTED until profile semantics, executable projection, and machine-checked safety invariants exist..

5.3 Confirmed difference

`confirmed-different` is symmetric for its subject pair but is not a general transitive relation. If a cannot link to b, and b cannot link to c, nothing follows automatically about a and c. An accepted partition must reject any cluster containing an active cannot-link pair and return a conflict witness rather than silently dropping the constraint.

5.4 Contradiction as data

A contradiction may arise from a proposed merge crossing a cannot-link, disagreement between authoritative sources, incompatible time intervals, type or cardinality violation, model/reviewer disagreement, or a source supersession that conflicts with local evidence. The system must preserve both sides and route the conflict according to policy. Last-writer-wins is not an identity semantics.

5.5 State transitions

Candidate transitions include:

$$\begin{aligned}
\texttt{candidate} &\to \texttt{possible}, \texttt{probable}, \texttt{confirmed-different}, \\
\texttt{possible} &\to \texttt{probable}, \texttt{confirmed-same}, \texttt{confirmed-different}, \\
\texttt{probable} &\to \texttt{confirmed-same}, \texttt{confirmed-different}.
\end{aligned}$$

Any active decision may become contradicted; any effective decision may be superseded; an admitted reversible decision may be retracted. Every transition records the actor or method, policy, time, authority check, and provenance. A rescore appends a new result; it does not mutate the old one.

6 Methods produce evidence, not identity

6.1 Deterministic rules

A deterministic rule may evaluate typed positive predicates p_i and negative predicates q_j:

$$\mathrm{Rule}_k(r_i, r_j) = \bigwedge_{\ell=1}^{n} p_\ell(r_i, r_j) \wedge \bigwedge_{m=1}^{s} \neg q_m(r_i, r_j).$$

Examples include equal source-certified immutable identifiers, equal normalized composite keys within a jurisdiction and interval, source-declared aliases, or an approved internal crosswalk.

The rule record must include normalizers, null behavior, key-reuse policy, time scope, negative predicates, and authority. Deterministic means reproducible under fixed input.

It does not mean true. Identifiers are mistyped and reused; household keys are confused with person keys; defaults compare equal; composite keys omit differentiating fields; source meaning changes. Every rule therefore needs adversarial fixtures and measured precision on its intended population.

Merge/purge and sorted-neighborhood work provides an established large-scale deduplication baseline [12]. Generic match/merge frameworks such as Swoosh expose useful algebraic conditions [2]. P7 must state whether any admitted operation actually satisfies idempotence, commutativity, associativity, or representativity; naming a typed rule does not grant those properties.

6.2 Probabilistic record linkage

For comparison vector γ, the classical likelihood ratio is

$$\Lambda(\gamma) = \frac{\Pr(\gamma \mid M)}{\Pr(\gamma \mid U)}, \qquad w(\gamma) = \log \Lambda(\gamma),$$

where M denotes matching pairs and U nonmatching pairs [8]. Thresholds divide pairs into link, possible-link/review, and nonlink actions under declared errors or costs. Operational record-linkage work demonstrates how such methodology is applied in practice [13].

P7 must not present w as a calibrated posterior by default, assume comparison-field independence, reuse thresholds across populations, ignore blocking-induced selection, or convert a high-probability pair into confirmed identity without policy and authority.

Every probabilistic result should retain the model and feature revisions, training or estimation population, blocking policy, prevalence if used, threshold and cost revision, calibration population, uncertainty, leakage audit, source/schema revisions, and result digest.

6.3 Calibration and decisions

If the interface says "70% match probability," roughly 70% of comparable held-out cases in the declared population should be matches. Required evidence includes reliability diagrams, Brier and log scores, calibration intercept and slope where applicable, bin counts, confidence intervals, and drift by source, time, entity type, and missingness. Proper scoring rules provide the relevant statistical foundation [9].

The action still depends on a declared loss

$$\mathcal{L}(a, y), \qquad a \in \{\text{link}, \text{review}, \text{nonlink}\}.$$

A high-consequence profile may require review at high probability. A calibrated probability is evidence for a policy decision, not an authority-bearing decision itself.

6.4 Candidate generation and vector retrieval

Candidate generation reduces the comparison space:

$$\mathsf{Candidates}_k(r) \subseteq R \setminus \{r\}.$$

Learned and iterative blocking are established approaches [18, 24]. Candidate methods are evaluated with pair completeness or candidate recall, reduction ratio, latency, and cost. A missed true pair is unavailable to a perfect downstream matcher.

Let $e : R \to \mathbb{R}^d$ be an embedding and d a distance. The relation

$$d(e(r_i), e(r_j)) \leq \epsilon$$

or a nearest-neighbor rank does not imply

$$\text{denotes}(r_i) = \text{denotes}(r_j).$$

The embedding is trained for a declared objective; proximity changes with model, preprocessing, metric, corpus, and threshold; approximate retrieval may miss items; similar entities can be distinct; one entity can drift apart; distance is generally nontransitive; rank depends on the indexed population; and no authority is encoded in the vector.

HNSW and Faiss establish scalable approximate similarity search [14, 17]. DeepMatcher and Ditto establish learned pair-classification baselines [16, 19]. None establishes semantic identity from raw proximity.

Requirement 6.1 (Vector safety boundary)**.** A vector adapter emits only a method result labelled `candidate`. It cannot invoke the effective-decision path.

Evidence status. OPEN. No adapter or misuse test exists..

6.5 Pair scores are not partitions

Suppose

$$\Pr(a = b) = 0.92, \qquad \Pr(b = c) = 0.91, \qquad \Pr(a = c) = 0.08.$$

Thresholding the first two scores and taking connected components merges all three records despite strong contrary evidence. A partition needs an explicit cluster policy over positive and negative evidence, constraints, time, objective, solver status, and deterministic tie-breaking.

Correlation clustering supplies a baseline objective for positive and negative pair labels [1]. Relationship-aware and collective resolution can improve decisions but makes evidence dependent [4, 6]. Declarative work on matching dependencies and LACE supplies rule, constraint, chase, and collective baselines [3, 5]. P7 cannot claim novelty merely for storing rules in a typed file.

6.6 Human review

Active learning, crowdsourcing, and end-to-end entity-matching systems already combine machine and human work [10, 15, 21, 23]. Human review is therefore neither a new contribution nor an oracle.

A review artifact should retain the evidence shown and withheld, reviewer and principal, authority scope, decision and confidence, reason, duration, exposure to model output, conflicts, second review, adjudication, and later reversal. The proposed CatDB delta is durable and reversible decision semantics across layers, not the existence of a review queue.

7 Authority, ownership, and provenance

7.1 Scoped authority

An authority grant is scoped by principal, organization, role, operation, entity type, source or canonical domain, tenant, jurisdiction, valid and system time, risk class, quorum, delegation chain, and revocation revision. Authority is checked when a decision becomes effective, not only when proposed.

A statistically reliable source may lack authority to decide a legal relationship. An authoritative source may still be wrong. A reviewer may understand the evidence but lack permission across a tenant boundary. These cases require separate fields and separate failure modes.

Evidence status. Authority non-bypass is OBSTRUCTED pending a shared policy semantics and adversarial implementation tests..

7.2 Source ownership

Canonicalization does not transfer source ownership. For every source record, the decision projection should disclose its owner, read/write policy, any governed attribute authority, and update route. For every canonical attribute, it should disclose whether the value was copied, selected, synthesized, or left disputed.

Entity identity and data fusion are distinct. Two records can be confirmed as the same entity while their addresses, names, balances, or dates remain in conflict. A merge that silently chooses one value has performed an additional, unrecorded decision.

7.3 Provenance projection

P7 depends on P13 for immutable source tokens, evidence derivation, method and policy identity, actor and delegation, revision, invalidation, downstream dependence, disclosure, and retention. P7 adds a reconciliation-specific decision object. Neither paper should invent incompatible artifact IDs.

One prospective W3C PROV projection is:

PROV-O establishes an interchange model for these relations [27]. It does not define P7's states, probability semantics, calibration, authorization, or confirmed equivalence. An exporter must therefore report information loss.

7.4 Audit questions

An implementation should answer:

1. Which source records belong to an anchor at a given valid and system time?
2. Which evidence supported and opposed the decision?
3. Which rule or model revision produced each result?
4. What population calibrated any displayed probability?

Table 3: Prospective P7 to PROV-O projection.

P7 artifact	PROV-O representation
source record or snapshot	`prov:Entity`, with specialization if needed
evidence observation	entity generated by an observation activity
model, rule, or policy	`prov:Plan` or versioned entity
matching or review	`prov:Activity`
reviewer, service, organization	`prov:Agent`
decision	entity generated by the activity
approval role	qualified association with `prov:Role`
delegation	qualified delegation
revised decision	`prov:wasRevisionOf`
retracted decision	qualified invalidation

5. Who proposed, approved, rejected, superseded, or reversed the decision?
6. Which authority and delegation were in force?
7. Who owns each source record?
8. Which mappings, queries, and materializations depend on the decision?
9. What became stale after contradiction or reversal?
10. Which evidence is hidden, expired, or redacted?
11. Can the prior projection be reproduced?

Every answer must state completeness and disclosure limits. Provenance describes origin and process; it does not prove that the identity decision is true.

8 Time, supersession, and reversible reconciliation

8.1 Bitemporal decisions

A decision may have valid time, describing when it is intended to hold in the modeled domain, and system time, describing when CatDB recorded it. This allows the statement "the records referred to the same legal organization during 2024" to remain separate from "the system learned in 2026 that its earlier conclusion was wrong."

The profile must state whether identity is endurant, time-indexed, or modeled through explicit succession. Adding timestamps without this policy does not resolve the conceptual choice.

8.2 Supersession

Supersession may mean corrected duplicate, account replacement, corporate succession, renamed identifier, source redirect, revised decision, or policy revision. Only some of these imply sameness. A supersession edge therefore needs a typed reason, direction, interval, authority, and declared effect on canonical membership.

8.3 Logical reversal

A merge is admitted as logically reversible only when:

1. every source record remains an immutable reference;
2. the prior canonical partition remains reconstructible;
3. the merge decision is append-only and can be superseded or retracted;
4. synthesized attributes retain derivation;
5. downstream artifacts name the decision revision they used;
6. reversal emits invalidation and recomputation obligations;
7. historical queries reproduce the earlier projection; and
8. failed reversal attempts remain recorded.

Logical reversal does not make physical repair free. Indexes and materializations may need rebuilding. External writes may need compensation. Deleted or redacted evidence may make reversal unsupported.

8.4 Splitting a merged cluster

Removing one pair edge is insufficient when a canonical cluster has many members. The system must compute a new partition, assign derived facts by provenance, retain ambiguous facts, and disclose external effects that cannot be undone. The first profile should support only a fragment whose affected decisions and dependencies are complete; other cases return `reconciliation-dependent` or `unsupported`.
Evidence status. Reversal and invalidation are OBSTRUCTED on P6, P9, and P13..

9 Cross-layer contracts

9.1 Composable integration

P4 mappings may transform source fields into typed candidate inputs. They must preserve original source references and mapping loss. They may not identify records merely because mapped fields align, treat object synthesis as canonical identity, or compose uncertain identity edges as lawful schema mappings.

Any $N + M$ integration-maintenance comparison must count identity profiles, exceptions, approvals, contradictions, and reversal work. If reconciliation effort grows pairwise, the broader maintenance claim may fail.

9.2 Federated query

A federated result that reconciles records must carry:

- identity profile and projection revision;
- source set, availability, and snapshot;
- candidate and matching methods;
- effective decision state;
- unresolved or contradictory records;
- confidence semantics and freshness; and
- provenance reference.

A strict query rejects when required identity decisions are unresolved. A partial query returns explicit unresolved groups and missing-source metadata. It must not silently drop duplicates or merge likely matches.

Evidence status. Identity-aware federation is OBSTRUCTED on the P5 result contract and runtime..

9.3 Provenance and materialization

P13 must not use identity evidence as proof of truth, and P7 must not use provenance as proof of identity. P6 must track every materialization or cache whose content depends on a decision revision. Reversal is sound only for the recorded dependency closure. An artifact outside that closure is a completeness failure, not harmless missing metadata.

9.4 Write propagation and workspaces

P8 cannot infer writeability from confirmed identity. A canonical anchor may unify reads while source owners retain distinct write authority. P9 cannot merge workspace decisions with a file-level union: merge, cannot-link, retraction, and authority changes can be noncommutative. Conflicts require semantic classification and, in some cases, coordination.

10 Formal core and theorem roadmap

10.1 Why the formal target is narrow

Formalization should verify the deterministic decision kernel, not whether a person, organization, or product match is true. A theorem may show that uncertain edges never enter the confirmed partition. It cannot show that the evidence justifying a confirmed edge is factually correct. A theorem may show authority derivations are present under a policy model. It does not verify authentication, credentials, deployment, or organizational behavior.

10.2 Finite core

The proposed finite model includes source references, profiles, evidence and method-result references, decision states, scoped grants, valid/system intervals, decision events, and normalized projections. Assumptions include decidable equality, finite histories, well-formed references, interval ordering, and acyclic supersession.
Evidence status. OBSTRUCTED until identity-profile v1 is frozen..

10.3 State-transition safety

Open conjecture 10.1 (Uncertain-state exclusion). For every well-formed profile and event history, adding, removing, or reordering events whose effective states are only `candidate`, `possible`, or `probable` leaves the confirmed-equivalence projection unchanged.

Open conjecture 10.2 (Decision witness). Every effective confirmed decision has a well-formed profile, actor, policy, authority derivation, evidence reference, and system-time record.

Both statements are OPEN CONJECTURE. No Lean representation or declaration exists.

10.4 Confirmed-equivalence projection

Open conjecture 10.3 (Projection soundness). For a well-formed finite event history without an active conflict witness, the projected `confirmed-same` relation is reflexive over admitted active references, symmetric, and transitive after explicit closure; every cluster member is traceable to effective confirmed decisions; and no active cannot-link pair occupies one cluster.

If closure conflicts with a cannot-link, the projection should be undefined with a conflict witness rather than silently repaired.

10.5 Replay

Open conjecture 10.4 (Historical replay). Under a deterministic total event order or a declared commutativity relation, repeated replay is idempotent, causally independent permutations are extensionally equal, system-time prefixes reproduce historical projections, and later events do not alter earlier system-time results.

Evidence status. OPEN CONJECTURE. No event semantics or implementation exists..

10.6 Restricted reversal

For an admitted merge m, candidate inverse decision m^{-1}, and state s, the desired law is

$$\mathsf{Project}\big(m^{-1}(m(s))\big) \approx \mathsf{Project}(s),$$

where extensional equality ignores new audit-event IDs but preserves source references, cluster membership, constraints, ownership, and prior facts whose derivation remains valid.

Open conjecture 10.5 (Restricted reversal law). The equation above holds for the explicitly admitted reversible fragment and fails closed when required provenance or dependency information is absent.

This statement cannot cover irreversible external writes or deleted derivations. It is OPEN CONJECTURE and operationally OBSTRUCTED.

11 Running example: reconciling customer records

11.1 Profile

Return to the three systems in the introduction. The profile asks:

> Do two source records describe the same natural person within tenant `north`, for customer-service analytics, during the overlapping valid interval?

This question excludes household, billing-account, and legal-party identity. The profile admits CRM, billing, and ticketing references; exact and normalized identifiers; incompatible birth date as negative evidence; a Fellegi–Sunter result; vector candidates; one tenant reviewer role; a source owner for each record; and a cannot-link constraint. Vector retrieval may propose a pair, but effective confirmed sameness requires human approval.

11.2 Source references

The immutable references are:

$$\begin{aligned}
r_c &= (\texttt{crm}, \texttt{customer}, 4821, s_c),\\
r_b &= (\texttt{billing}, \texttt{account}, 773, s_b),\\
r_t &= (\texttt{ticketing}, \texttt{contact}, 51, s_t).
\end{aligned}$$

The billing key names an account, not necessarily a person. Its mapping into this profile therefore requires a typed person-holder projection and may be partial. The lexical key `773` itself supplies no cross-source identity.

11.3 Candidate and evidence

An embedding index returns (r_c, r_b) at rank 1 and (r_c, r_t) at rank 3. Each result records the model, projected fields, index revision, metric, search parameters, rank, score, source filters, time, and an exact-search recall audit. Both remain `candidate`; neither changes an anchor.

For (r_c, r_b), the evidence includes an equal verified email, compatible name, overlapping interval, and a billing-holder identifier whose source regime is documented. For (r_c, r_t), it includes a similar name and address but incompatible authoritative birth dates. Missing phone number is recorded as missing, not as negative evidence.

11.4 Method results and decision

A versioned linkage method produces $w(\gamma) = 4.2$ for (r_c, r_b), in the review region, and $w(\gamma) = 3.9$ for (r_c, r_t). These values are likelihood-ratio scores, not advertised as probabilities. A supervised classifier separately reports 0.88 and 0.84; because its output is not calibrated on this tenant and source mixture, the profile labels the values `classifier-score`, not `match-probability`.

The tenant reviewer examines the first pair and approves `confirmed-same` under their authority for customer-service person identity. The second pair becomes `contradicted` because positive similarity conflicts with the birth-date cannot-link. The current projection creates one anchor for r_c and r_b, keeps r_t separate, and preserves all evidence.

11.5 Attribute conflict

The merged records have different postal addresses, and an identity decision does not choose between them. A canonical query may return both with provenance, apply a separate valid-time and attribute-authority rule, or return an explicit conflict. Silently copying the billing address would conflate identity with fusion.

11.6 Reversal

Later, the billing source reports that account 773 was reassigned. This assertion is new evidence, not a mutation of the original record. The prior decision becomes contradicted, then an authorized reviewer retracts it. A supported reversal would:

1. reconstruct the prior partition;
2. preserve the original merge and retraction events;
3. invalidate the reconciled customer materialization;
4. disclose affected historical query ranges;
5. recompute derived analytics; and
6. report any external write requiring manual compensation.

This sequence is illustrative. The repository has no executable profile, score, decision, projection, or reversal corresponding to it.

12 Prior art and novelty boundary

12.1 Classical linkage and duplicate detection

Fellegi and Sunter provide the canonical decision-theoretic record-linkage framework; Jaro develops and applies operational methodology [8, 13]. Merge/purge addresses scalable duplicate detection [12]. Swoosh studies generic match/merge functions and properties [2]. P7's states, authority, and bitemporal history are not results of those papers, and P7 cannot claim their underlying algorithms as new.

12.2 Uncertain, collective, and declarative resolution

Identity-uncertainty work models uncertain mappings from observations to entities [20]. Reference reconciliation and collective entity resolution use relational context and positive/negative evidence [4, 6]. Correlation clustering provides a partitioning baseline [1]. Matching dependencies and LACE show that declarative rules, matching functions, hard constraints, soft rules, and resolved answers predate P7 [3, 5].

These sources set a high bar. A typed syntax for rules or a graph of pair scores is not a contribution by itself. P7 must demonstrate an operational cross-layer property that established ER plus ordinary event logging and policy do not already supply.

12.3 Blocking, learning, and human workflow

Learned and iterative blocking are established candidate techniques [18, 24]. Interactive deduplication, CrowdER, and Corleone establish active and crowd review workflows [10, 21, 23]. Magellan establishes a complete entity-matching management baseline [15]. DeepMatcher and Ditto establish neural and pretrained-language-model pair classifiers [16, 19].

Therefore, P7 must compare setup, candidate generation, matching, review, correction, audit, and maintenance costs. Accuracy for one classifier is not evidence that the full CatDB workflow is better.

12.4 Vector search

HNSW and GPU similarity systems establish approximate search at scale [14, 17]. Their role in P7 is retrieval. The safety claim that a vector result cannot directly invoke a decision path could be a useful interface property, but the conceptual distinction between retrieval and classification is not new.

12.5 Semantic Web identity

OWL's `owl:sameAs` denotes equality of individuals [26]. Empirical and conceptual work documents the risks of using it for weaker relations [11]. SKOS distinguishes weaker mapping relations such as `closeMatch` and `exactMatch` for concepts [25]. A P7 exporter may use `owl:sameAs` only when the confirmed profile warrants substitutability. Candidate, probable, contextual, and concept-mapping relations need weaker, explicit predicates.

12.6 Provenance, trust, and privacy

PROV-O supplies normative provenance structures [27]. TruthFinder and source-dependence work study reliability, copying, and conflicting claims [7, 28]. Proper scoring rules support probabilistic evaluation [9]. Bloom-filter linkage supplies one privacy-preserving record-linkage approach [22].

P7 must not redescribe these results as CatDB innovations. A privacy mode requires an attacker and leakage model; provenance requires a loss audit; reliability remains distinct from authority.

12.7 Plausible CatDB delta

The strongest plausible novelty is a versioned identity assertion and decision layer that:

- retains source-local references and uncertainty;
- separates evidence, scores, authority, and decisions;
- exposes identity profile and time in every projection;
- supports an admitted, provenance-complete reversal fragment;
- invalidates mapped, federated, materialized, and workspace artifacts;
- prevents vector candidates from becoming identity by interface; and
- makes conflicting decisions reviewable rather than order-dependent.

This combination is an ENGINEERING HYPOTHESIS. “Category-aware identity semantics” is SPECULATIVE until a result unavailable to ordinary typed graphs, rules, and event histories is precisely stated and established.

13 Implementation roadmap

13.1 Profile and fixtures

The first milestone freezes identity-profile v1, its validation schema, diagnostic registry, normalization, digest profile, P4/P5/P13 reference contract, privacy boundary, and unsupported list. Exit requires agreement that every state and field has one meaning, candidate and probable never mean equality, and no source reference can be lost by design.
Evidence status. OBSTRUCTED on P4 and P13..

The second milestone builds hand-worked positive and negative fixtures for source identity, deterministic rules, probabilistic scoring, learned matcher metadata, vector candidates, clustering, authority, contradiction, time, reversal, federation, provenance, and privacy. Expected results must be independently derived rather than emitted by an implementation.
Evidence status. OPEN..

13.2 Independent Haskell reference

The proposed reference modules are:

```
CatDB.Identity.Types
CatDB.Identity.Validate
CatDB.Identity.Rules
CatDB.Identity.FellegiSunter
CatDB.Identity.Project
CatDB.Identity.Reversal
CatDB.Identity.Provenance
```

Properties include conservation of source references, exclusion of uncertain states from confirmed partitions, satisfaction of cannot-links, replay invariance for causally independent events, admitted reversal, acyclic supersession, and idempotent normalization.
Evidence status. OPEN. No module exists..

13.3 Rust production crate

The proposed `catdb-identity` crate owns production IR, validation, deterministic rule evaluation, candidate/result ingestion, policy checks, decision append, projection, audit queries, reversal planning, provenance hooks, and stable diagnostics. It should not own vector indexing, model training, authentication, or a general policy engine.
Evidence status. OPEN. No crate exists..

13.4 Method adapters and review workflow

Candidate adapters should cover exact blocking, one learned blocker, exact vector search, and HNSW. Matcher adapters should cover Fellegi–Sunter and one supervised pair classifier. A human-review adapter ingests a typed review artifact. Every adapter emits method results, never effective decisions.

Authority implementation includes scoped grants, delegation, proposal, approval, rejection, contradiction, supersession, retraction, quorum, and separation of duty. Authentication and the policy engine remain external dependencies.
Evidence status. Adapters are OPEN; authority workflow is OBSTRUCTED..

13.5 Projection, reversal, and integration

Later milestones add bitemporal projection, constraint checks, deterministic replay, supported reversal/split, downstream invalidation, strict and partial federation, workspace-local proposals, semantic decision diffs, and historical reproduction.
Evidence status. OBSTRUCTED on P5, P6, P9, and P13..

13.6 Security and privacy

Only after threat-model review should the system add evidence access control, sensitive-feature redaction, audit-safe disclosure, encrypted or tokenized linkage, key rotation, retention handling, and inference tests. A type named `PrivateIdentity` is not a privacy proof.
Evidence status. OPEN for implementation; formal noninterference remains SPECULATIVE..

14 Experimental evaluation contract

14.1 Research questions

The evaluation asks whether typed artifact separation prevents score-to-identity promotion; whether Rust and Haskell agree; whether probabilities are calibrated; how candidate methods trade recall and cost; how pairwise, collective, and constraint-aware policies trade error; what human review costs; whether provenance reduces audit and repair time; whether admitted reversal reconstructs prior partitions and invalidates all recorded dependents; what overhead full history adds; and whether any maintenance benefit survives identity exceptions and approval work.

14.2 Compared systems and ablations

The comparison matrix includes deterministic rules, Fellegi–Sunter, supervised and transformer pair classifiers, vector-candidate-plus-classifier, connected components as an intentionally unsafe baseline, correlation clustering, collective ER, human-only review, hybrid review, and the proposed typed/versioned CatDB workflow. Ablations remove provenance, authority checks, and reversible history.

Baseline implementations must be executable or clearly labelled faithful reconstructions. A toy baseline cannot establish superiority over a mature system.

14.3 Datasets and ground truth

Use hand-checkable synthetic data, corruption generators, licensed public entity-matching benchmarks, relational data for collective resolution, temporal source corrections, source-copying conflicts, domain shift, and a consent-safe enterprise-like dataset with independent adjudication. Every dataset manifest records provenance, license, unit of identity, ground-truth method, ambiguous labels, split, contamination, and checksums.

Sensitive-person linkage is excluded unless governance, privacy, licensing, and ethics review explicitly admit it.

14.4 Correctness metrics

Report candidate recall and reduction before matcher accuracy. Report pair precision, recall, false-match rate, and missed-match rate with denominators. Report cluster metrics with exact definitions, exact partitions for small cases, must-link/cannot-link violations, contradiction detection, source-reference conservation, replay determinism, reversal equivalence, invalidation recall, historical reproduction, misuse rejection, and interchange loss.

A single F1 score is insufficient because it hides prevalence, candidate loss, cluster shape, calibration, review burden, and consequences.

14.5 Probabilistic and systems metrics

Probabilistic evaluation includes Brier score, log loss, calibration intercept/slope, reliability curves, expected calibration error with bins, area under precision–recall, threshold precision/recall, review-region size, expected decision cost, intervals, and drift strata.

Systems evaluation includes profile compilation, candidate generation, exact/approximate recall and latency, scoring, policy evaluation, projection, replay, historical query, review throughput, storage amplification, provenance query, reversal, invalidation, recomputation, source rows/bytes, memory, CPU, privacy overhead, and failed/conflicted operations.

14.6 Human review and reversal studies

The human study preregisters recruitment, task, evidence presentation, model-score visibility, randomization, blinding, adjudication, training, disagreement, stopping, compensation, privacy/ethics approval, endpoints, and exclusions. It measures accuracy, time, disagreement, calibration, correction, reversal, and auditability.

The reversal study seeds false pair merges, false transitive clusters, late cannot-links, source corrections, supersession, model recalibration, revoked approval, mapping revision, newly visible evidence, and deletion. It measures partition recovery, fact reassignment, invalidation completeness, repair time, and external compensation.

14.7 Promotion gates

Before confirmatory execution, freeze source state, dependencies, profile, fixtures, datasets, splits, models, thresholds, policies, baselines, hypotheses, margins, sample-size plan, seeds, environment, and exclusions. Preserve failed runs.

Examples of promotion criteria are:

- zero source-reference loss in the supported fixture family;
- exact Rust/Haskell action agreement under a frozen numeric profile;
- calibration within preregistered held-out margins;
- candidate-recall lower bound above the declared threshold;
- zero authority bypass in the admitted adversarial suite;
- byte-identical normalized replay;
- exact extensional recovery for every admitted reversal;
- zero missed recorded downstream dependency;
- zero unqualified partial-source result; and
- preregistered audit/repair improvement with noninferior error.

None of these criteria has been run.

15 Current results and implementation status

15.1 What this paper establishes

This author pass establishes only a specification-level result: the proposed claim has been decomposed into artifact types, states, authority boundaries, cross-layer obligations, formal conjectures, falsifiers, and experiments. The paper's prior-art claims are SUPPORTED BY PRIOR LITERATURE where cited. The repository audit supports the statement that P7 implementation and experiment artifacts are absent at this evidence cut.

15.2 What this paper does not establish

Table 4: Current P7 status by layer.

Layer	Status	Reason
record-linkage prior art	SUPPORTED BY PRIOR LITERATURE	Primary sources establish the field.
vector/identity boundary	SUPPORTED BY PRIOR LITERATURE plus inference	ANN defines retrieval; OWL defines strong equality.
identity profile	OBSTRUCTED	P4/P13 references are not frozen.
typed artifact model	OPEN	Manuscript and dossier only.
deterministic/probabilistic methods	OPEN	No implementation or fixtures.
authority and ownership	OBSTRUCTED	Shared policy semantics absent.
Rust runtime	OPEN	Crate absent.
Haskell oracle	OPEN	Modules absent.
Lean laws	OPEN	Representation absent.
federation	OBSTRUCTED	P5 result/runtime contract absent.
reversal and invalidation	OBSTRUCTED	P6/P9/P13 dependencies absent.
privacy/security	OPEN	Threat model and implementation absent.
experiments	OPEN	No run artifacts.

No CatDB accuracy, calibration, scalability, review-quality, reversibility, privacy, or maintenance claim is experimentally supported. No P7 theorem is machine-checked. No existing categorical identity-mapping fixture is entity identity evidence.

16 Limitations

16.1 Conceptual limits

An identity profile makes the question explicit, but it cannot make the answer objective. A profile may be poorly scoped, biased, internally inconsistent, or politically contested. Canonicalization can still be useful without being universally true. Legal, ethical, and institutional constraints do not reduce to graph closure.

16.2 Statistical limits

Probabilistic methods depend on the population, sampling, comparison features, blocking, prevalence, calibration, and drift. Ground truth may itself be ambiguous or wrong. A reported probability can be well calibrated overall and still fail for a consequential subgroup. Collective inference may double-count copied evidence or amplify an early false merge.

16.3 Systems limits

Versioning and provenance cost storage, latency, review time, and operational work. Reversal may require expensive recomputation, while some external effects cannot be reversed at all. Partial source access limits completeness. A typed record design does not provide authentication, authorization enforcement, encryption, key management, or side-channel resistance.

16.4 Formal limits

The proposed theorems concern well-formed finite histories and policy derivations. They cannot establish that observations are correct, reviewers are honest, a statistical model transports, or a real-world entity exists. Even a future Lean proof of projection laws would not validate the empirical identity evidence.

16.5 Evaluation limits

Synthetic ground truth may miss the ambiguity found in production. Public benchmarks may be contaminated by pretrained models or encode only one identity question. Human studies depend on participant expertise, evidence presentation, and incentives. A positive result would support only the tested profile, workload, sources, and population.

17 Open questions

1. Which identity questions and entity types belong in profile v1?
2. Should confirmed identity be time-indexed, or should change use succession relations?
3. Which source-key guarantees permit deterministic promotion?
4. How should copied and syndicated evidence be represented?
5. Which numeric profile gives reproducible linkage scores across languages?
6. Which probability semantics may be shown to users?
7. How should calibration be corrected after blocking or population shift?
8. Which uncertain states may downstream queries consume?
9. How should pair probabilities become cluster proposals?
10. Which constraints are hard, defeasible, or authority-overridable?
11. What conflict witness is minimal and useful?
12. When may one source owner authorize identity across another owner's record?
13. Which decisions require quorum or separation of duty?
14. How should identity and attribute-fusion policies compose?

15. Which merge fragment is exactly reversible?
16. How should synthesized facts be split after an erroneous merge?
17. Which external effects require compensation?
18. What downstream invalidation completeness can P6 and P13 provide?
19. Which reconciliation operations commute across workspaces?
20. What may a redacted provenance answer reveal through graph shape?
21. Which privacy-preserving protocol and attacker model fit v1?
22. How should deletion interact with audit and reproducibility?
23. Which datasets have defensible and licensed ground truth?
24. How should disputed benchmark labels be scored?
25. Can the proposed layer demonstrate value beyond established ER, provenance, event histories, and ordinary policy systems?

18 Conclusion

Identity integration cannot be reduced to choosing the nearest vector or the largest score. It requires preserving source records and evidence, stating the exact identity question, applying a versioned method, obtaining authority under a declared policy, and exposing the consequences of the decision.

This paper proposes that CatDB make those objects first-class. It keeps source-local references intact and excludes candidate, possible, and probable states from confirmed equivalence. Contradictions survive approval. Canonical anchors are relative to a profile and time. Provenance records origin and process without claiming truth, and reversal is admitted only when the prior partitions and downstream effects can be reconstructed.

The proposal is not yet a CatDB capability. Its profile, runtime, fixtures, formalization, security model, and experiments are absent or blocked by upstream semantics. Vector similarity may generate candidates, but it cannot establish semantic identity by itself. The research program succeeds only if implementation and experiments show that typed, versioned, authority-bearing decisions materially improve audit and repair without hiding errors or costs. Until those gates close, the correct status is ENGINEERING HYPOTHESIS, OPEN, and OBSTRUCTED, not solved identity.

A Exact nonclaims

This manuscript makes the following exclusions binding:

1. CatDB does not invent record linkage, entity matching, entity resolution, blocking, merge/purge, collective ER, or human review.

2. Category theory does not discover real-world identity.
3. A schema mapping does not show that two records denote one entity.
4. A source identifier is not automatically global, stable, true, or unrecycled.
5. A canonical anchor is a profile-relative decision artifact, not proof of metaphysical or legal identity.
6. Candidate, possible, and probable relations are not equivalence relations.
7. Similarity is evidence, not identity.
8. Vector similarity and nearest-neighbor rank may generate candidates but cannot establish semantic identity by themselves.
9. Approximate vector search may miss candidates.
10. A high classifier score is not a calibrated probability unless validated as such.
11. A calibrated probability is not an authority-bearing decision.
12. Deterministic rules and human reviewers can be wrong.
13. More agreeing sources do not imply independent evidence.
14. Source reliability is not source authority.
15. Source authority is scope- and time-relative.
16. Confirmed identity does not resolve attribute conflicts.
17. Supersession does not automatically imply sameness.
18. Connected components over scores are not a safe universal cluster semantics.
19. `owl:sameAs` is not a fuzzy or candidate relation.
20. W3C PROV does not supply P7 identity semantics.
21. Provenance does not prove identity or truth.
22. Append-only history alone does not guarantee authenticity, completeness, or reversibility.
23. Logical reversal does not remove physical repair or compensation.
24. Reversal may be unsupported after evidence deletion or redaction.
25. Identity decisions do not generally commute under replication.
26. Partial source availability is not a complete identity view.
27. Privacy-preserving linkage is not private against every attacker.
28. Benchmark accuracy does not establish production quality.
29. F1 does not measure calibration, candidate recall, cluster quality, review burden, or decision cost.

30. CatDB does not eliminate manual reconciliation, ETL, data movement, source governance, or distributed-systems tradeoffs.
31. Existing schema-path and mapping work does not establish P7.
32. A future formal projection proof would not prove match evidence.
33. Future Rust/Haskell agreement would be computational conformance, not proof of identity truth.

B Minimum fixture matrix

Table 5: Minimum hand-worked fixture families before P7 claim promotion.

Family	Positive coverage	Adversarial coverage
Source identity	stable key, namespace separation, tombstone, alias evidence	key reuse, household/person type mismatch, cross-time ambiguity
Deterministic rules	authoritative identifier, normalized composite key, permutation stability	null/default collision, schema meaning change, rule-priority conflict
Probabilistic linkage	hand score, match/review/nonmatch regions, threshold revision	zero probability, missing comparisons, prior shift, blocking selection
Learned matcher	model/feature digests, calibrated output, threshold revision	uncalibrated probability label, leakage, drift, missingness failure
Vector candidates	exact hit, HNSW result plus exact audit, candidate recall	true-pair miss, high-similarity cannot-link, low-similarity true match
Clusters	consistent confirmed triangle, deterministic tie break	nontransitive scores, must/cannot conflict, false collective propagation
Authority	in-scope owner, valid delegation, quorum, dual approval	out-of-scope action, revoked delegation, model self-approval
Contradiction	positive and negative evidence coexist, reviewer disagreement	last-writer-wins attempt, hidden negative evidence, interval conflict
Time and reversal	historical cut, merge/reversal, supported split, invalidation	incomplete provenance, ambiguous derived fact, external compensation

Family	Positive coverage	Adversarial coverage
Cross-layer	mapping creates candidate, strict/partial federation, PROV trace	lost source ref, silent partial source, stale cache, `owl:sameAs` misuse
Privacy	declared raw/token/Bloom profiles, key rotation	frequency attack, linkage leakage, unsupported privacy claim

C Claim-promotion ledger

Table 6: Selected P7 claim anchors and minimum promotion gates.

ID	Claim	Status	Promotion gate
`ID-01`	Source references survive canonicalization and reversal.	OPEN	Shared fixtures and exact Rust/Haskell agreement.
`ID-02`	Candidate, evidence, method result, and decision remain typed separately.	OPEN	Schema, validators, and negative fixtures.
`ID-03`	Uncertain states never enter confirmed equivalence.	OBSTRUCTED	Frozen state semantics, executable projection, Lean invariant.
`ID-04`	Confirmed-same projects to an equivalence relation within one profile/time cut.	OBSTRUCTED	Precise projection and machine-checked laws.
`ID-05`	Accepted projections respect confirmed-different.	OBSTRUCTED	Constraint semantics, exhaustive finite tests, Lean invariant.
`ID-07`	Rust and Haskell agree on reference linkage scores.	OPEN	Hand-worked scores and frozen numeric tolerance.
`ID-08`	Displayed probabilities are calibrated.	OPEN	Held-out evaluation with proper scores and intervals.
`ID-09`	Vector search emits candidates only.	OPEN	API restrictions and misuse tests.
`ID-12`	Scoped authority is non-bypassable.	OBSTRUCTED	Policy IR and adversarial API/backend tests.
`ID-15`	Replay is deterministic at a system-time cut.	OPEN	Event fixtures, shuffled delivery, byte-stable projection.

ID	Claim	Status	Promotion gate
`ID-16`	Admitted reversal reconstructs the prior partition.	OBSTRUCTED	Reversible fragment, Rust/Haskell agreement, Lean law.
`ID-17`	Reversal invalidates every recorded dependent artifact.	OBSTRUCTED	P6/P13 dependency graph and mutation tests.
`ID-18`	Contradictions are preserved without hidden order.	OPEN	Permutation and concurrency fixtures.
`ID-19`	OWL export reserves equality for qualifying confirmed identity.	OPEN	Export profile and loss/misuse tests.
`ID-21`	Partial source availability is always qualified.	OBSTRUCTED	P5 result envelope and source-failure fixtures.
`ID-24`	Privacy mode meets a declared leakage contract.	OPEN	Independent security analysis and attacks.
`ID-25`	The layer reduces audit/repair cost without degrading error.	OPEN	Preregistered full-workflow comparison.

References

[1] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56: 89–113, 2004. doi: 10.1023/B:MACH.0000033116.57574.95.

[2] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: A generic approach to entity resolution. *The VLDB Journal*, 18 (1):255–276, 2009. doi: 10.1007/s00778-008-0098-x.

[3] Leopoldo Bertossi, Solmaz Kolahi, and Laks V. S. Lakshmanan. Data cleaning and query answering with matching dependencies and matching functions. In *Proceedings of the 14th International Conference on Database Theory*, 2011. URL https://openproceedings.org/2011/conf/icdt/BertossiKL11.pdf. Extended journal version DOI 10.1007/s00224-012-9402-7.

[4] Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007. doi: 10.1145/1217299.1217304.

[5] Meghyn Bienvenu, Gianluca Cima, and Víctor Gutiérrez-Basulto. LACE: A logical approach to collective entity resolution. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2022. doi: 10.1145/3517804.3526233.

[6] Xin Dong, Alon Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 85–96, 2005. doi: 10.1145/1066157.1066168.

[7] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009. doi: 10.14778/1687627.1687690.

[8] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

[9] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

[10] Chaitanya Gokhale et al. Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014. doi: 10.1145/2588555.2588576.

[11] Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. When `owl:sameAs` isn't the same: An analysis of identity in linked data. In *The Semantic Web – ISWC 2010*, 2010. doi: 10.1007/978-3-642-17746-0_20.

[12] Mauricio A. Hernández and Salvatore J. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 127–138, 1995. doi: 10.1145/223784.223807.

[13] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. doi: 10.2307/2289924.

[14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

[15] Pradap Konda et al. Magellan: Toward building entity matching management systems. *Proceedings of the VLDB Endowment*, 9(12):1197–1208, 2016. doi: 10.14778/2994509.2994535.

[16] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60, 2020. doi: 10.14778/3421424.3421431.

[17] Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

[18] Matthew Michelson and Craig A. Knoblock. Learning blocking schemes for record linkage. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, 2006. URL https://aaai.org/Papers/AAAI/2006/AAAI06-070.pdf.

[19] Sidharth Mudgal et al. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*, 2018. doi: 10.1145/3183713.3196926.

[20] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15*, 2002. URL https://papers.nips.cc/paper_files/paper/2002/hash/d30960ce77e83d896503d43ba249caf7-Abstract.html.

[21] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002. doi: 10.1145/775047.775087.

[22] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 9(41), 2009. doi: 10.1186/1472-6947-9-41.

[23] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. CrowdER: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 5(11):1483–1494, 2012. doi: 10.14778/2350229.2350263.

[24] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, 2009. doi: 10.1145/1559845.1559870.

[25] World Wide Web Consortium. SKOS simple knowledge organization system reference. https://www.w3.org/TR/skos-reference/, 2009. W3C Recommendation.

[26] World Wide Web Consortium. OWL 2 web ontology language: RDF-based semantics, second edition. https://www.w3.org/TR/owl2-rdf-based-semantics/, 2012. W3C Recommendation.

[27] World Wide Web Consortium. PROV-O: The PROV ontology. https://www.w3.org/TR/prov-o/, 2013. W3C Recommendation.

[28] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007. doi: 10.1145/1281192.1281309.