Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agreement is not corroboration: bounding the independence of cultural-heritage authority links

Emin Talip Demirkiran

etd@eskisehir.edu.tr

Eskisehir Technical University https://orcid.org/0000-0001-6063-8178

Research Article

Keywords: Wikidata, authority control, data provenance, source dependence, partial identification, digital libraries

Posted Date: September 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11053378/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Purpose.

Agreement across authority records supports reconciliation, but it does not by itself establish independent corroboration when one identifier may have been imported from another source or both may share an upstream origin. This study asks how much provenance independence can be established within Wikidata-mediated concordance between Getty ULAN and Library of Congress (LC) name-authority identifiers.

Methods.

A pre-specified analysis examined 1,546 statement-instances from a frozen 3,000-entity analysis pool. Statement provenance was classified as PROPAGATED, INDEPENDENT, or UNTRACEABLE. Because unresolved derivation cannot be assigned without additional assumptions, the overall independence rate was treated as a partially identified parameter and estimated under hybrid post-stratification with entity-cluster bootstrap uncertainty.

Results.

Of 1,546 statements, 728 were PROPAGATED, 256 INDEPENDENT, and 562 (36.4%) UNTRACEABLE. Among traceable statements, approximately 71% were propagated. The post-stratified independence rate was identified only within [0.163, 0.556], spanning the substantive threshold of 0.50. Corrective analyses for a temporal-variable discrepancy, alternative weighting, and stratum-preserving bootstrap resampling did not change that identification conclusion. The analysed LC-name pathway was concentrated almost entirely in ULAN-linked persons.

Conclusion.

Cross-authority agreement can improve interoperability without constituting independent evidence. In digital-library and linked-data workflows, provenance dependence should be modelled explicitly because missing provenance can determine identifiability rather than merely reduce metadata completeness.

1 Introduction

Authority control increasingly crosses institutional and technical boundaries. In digital libraries, authority links support metadata integration, entity resolution, cross-collection navigation, discovery, and reuse across heterogeneous collections. Libraries, archives, museums, and digital-humanities projects routinely connect local records to external authority files and knowledge graphs through identifiers and reconciliation. Wikidata has become an important part of this infrastructure: studies in libraries and

GLAM institutions describe its use for linking, disambiguation, enrichment, curation, and metadata exchange [1–4]. More recent work documents Wikidata-mediated reconciliation in operational authority-file, museum-record, and repository workflows [5–8].

Reconciliation, however, answers a different question from corroboration. Two systems may contain the same identifier because they were curated independently, because one copied or imported the identifier from the other, or because both inherited it from a common upstream source. These situations are observationally similar at the level of value agreement but evidentially different. The distinction is familiar in the source-dependence and truth-discovery literature, where repeated claims cannot be treated as independent votes when sources copy from one another [9–11].

Cultural-heritage and Wikidata research has addressed closely related problems. Prior work has examined ambiguity and inconsistency in cross-authority reconciliation, semantic relations in vocabulary mappings, provenance-rich interlink creation, and the quality or completeness of Wikidata references [12–16]. These studies establish why provenance matters, but they do not directly quantify derivational independence within authority links that already agree. Agreement can therefore appear most reassuring precisely where dependence is least visible.

The study addresses one bounded research question:

RQ. Within the analyzable Wikidata-mediated ULAN–Library of Congress name-authority concordance, how much of the observed agreement can be established as independently derived when statement provenance is incomplete?

The design retains missing derivation evidence as UNTRACEABLE rather than forcing it into either PROPAGATED or INDEPENDENT. This choice changes the inferential target. When a substantial share of provenance cannot be resolved, the relevant independence rate is not point identified; instead, the observed data support an interval of values corresponding to the unresolved latent states.

The paper makes three contributions. First, it operationalises source dependence in an authority-control setting as a statement-level partial-identification problem, separating what observed provenance establishes from what would require assumptions about missing derivation evidence. Second, it provides an empirical decomposition of traceable agreement in the estimable ULAN–LC-name slice, distinguishing propagated from independently established assertions. Third, it documents the effective scope of the concordance mechanism itself, showing that the analysed LC-name pathway is substantially narrower than a generic “Getty Vocabularies” framing would imply.

The contribution is therefore not that reconciliation, provenance, or source dependence are individually new topics. It lies in combining them to ask whether already-concordant authority links constitute independent evidence and, when provenance is incomplete, reporting the resulting identified set rather than replacing missing derivation with an untestable point assumption. Within the targeted literature searches conducted for this study, no directly overlapping empirical study was identified with this

combination of domain, unit of analysis, and partial-identification treatment for Wikidata-mediated ULAN–LC name-authority links. This search-bounded statement is not presented as an absolute priority claim.

2 Related Work

2.1 Wikidata and authority reconciliation

Wikidata's role in authority control is supported by both reviews and applied studies. A systematic review of library use describes Wikidata as an open structured-data platform for linking local and global metadata and as a candidate infrastructure for authority-data disambiguation and linking [1]. Bianchini, Bargioni, and Pellizzari di San Girolamo compare Wikidata and VIAF specifically in library authority control and argue for complementary roles [2]. Reviews of digital-humanities and GLAM practice likewise document Wikidata-based enrichment, curation, reuse, and cross-collection linking [3, 4].

Applied work makes this reconciliation role more concrete. Pellizzari di San Girolamo traces a decade of SBN–Wikidata authority-record reconciliation and its practical difficulties [5], and later examines reciprocal cooperation between Wikidata and authority files more broadly [8]. Trognitz and colleagues describe Wikidata as a reconciliation anchor for identifying missing identifiers and controlled vocabularies in long-term repository workflows [7]. Together, these studies establish the operational importance of cross-authority linking, but they do not estimate whether the resulting agreements are derivationally independent.

2.2 Reconciliation ambiguity and semantic mapping

A related literature asks whether cross-system mappings are correct, usable, or semantically appropriate. Daquino, Mambelli, and Kozlov reconcile art-historical entities across LOC, GND, VIAF, ULAN, and Wikidata, highlighting substantial inconsistency among cross-reconciled URIs and the need to account for institutional cataloguing practices and human validation [12]. Foster, Norling, and Westerby similarly show that reconciling museum constituent records to Wikidata can expose incomplete identifiers and conflicts between local and external records [6].

At vocabulary level, Zhang and Yan map UNESCO-inscribed Chinese intangible-cultural-heritage elements across Wikidata, LCSH, and Getty AAT using explicit mapping relations rather than treating every cross-system association as exact identity [13]. These studies are close empirical neighbours because they operate across overlapping authority and vocabulary infrastructures. Their target, however, is mapping quality or semantic alignment. The present study begins after concordance has already been established and asks whether the concordant assertions were derived independently.

2.3 Provenance, reference quality, and uncertainty

Provenance-rich interlinking provides another neighbouring literature. McKenna, Debruyne, and O'Sullivan's NAISC-L framework creates metadata interlinks while recording who created a link, where

and when it was created, how it was produced, and why it was asserted [14]. This work makes explicit that the evidential history of an interlink is not merely an implementation detail. It is prospective, however: the framework is designed to preserve provenance at link creation.

Recent digital-library work similarly models provenance for multi-source knowledge assertions. Menotti and colleagues extend nanopublications with a knowledge-provenance graph that represents how assertions aggregate supporting and conflicting evidence from multiple sources [17]. That work addresses prospective representation and serialization of source lineage and trust. The present study addresses a complementary retrospective problem: when already-concordant authority assertions exist but their derivation history is incomplete, how much independence can the observed provenance establish?

The present setting is retrospective. The analysis audits already-existing assertions for which provenance may be incomplete. Recent Wikidata quality research reinforces the relevance of this distinction. RQSS evaluates completeness, verifiability, and other quality dimensions of Wikidata references [15]. Di Pasquale and colleagues show that uncertain or otherwise weaker-logical-status information in cultural-heritage records can be sparsely and ambiguously represented [16]. Broader Wikidata quality work likewise documents the difficulty of assessing statement quality in a collaboratively maintained knowledge graph [18]. These studies motivate explicit treatment of missing or ambiguous provenance, but they do not recover the latent derivation of statements whose provenance is absent.

2.4 Source dependence and the gap addressed here

The theoretical distinction between agreement and independent corroboration is explicit in source-dependence research. Dong, Berti-Équille, and Srivastava show that copying relationships among data sources can distort truth discovery because repeated claims cease to provide independent support [9]. Subsequent surveys treat source reliability and dependence as core components of multi-source inference [10], while recent work continues to model dynamic and complex dependence rather than assume source independence [11].

Most truth-discovery methods infer dependence statistically from many claims and sources, often where values conflict. The present setting differs in two respects. First, the empirical problem is agreement rather than conflict: the ULAN and LC identifiers already concord. Second, Wikidata exposes statement-level provenance cues that permit some assertions to be classified directly as propagated or independently established. The cost of this specificity is an unresolved UNTRACEABLE class whose latent derivation cannot be learned from the observed provenance alone.

The resulting gap lies at the intersection of reconciliation, provenance, and source dependence. Existing reconciliation studies ask whether entities or identifiers align and where those alignments fail; provenance frameworks ask how link history should be represented; Wikidata quality studies ask whether statements and references are adequately documented; and truth-discovery methods model

dependence in generic multi-source settings. The present study instead asks whether already-concordant Wikidata-mediated ULAN–LC name-authority assertions are independently established, propagated, or untraceable, and what can be identified about the independence rate when the third class remains unresolved.

3 Data and Scope

3.1 Data sources and frozen analysis frame

The study uses Wikidata (CC0), Getty Vocabularies (ODC-By), and Library of Congress name-authority data (public domain). VIAF was initially considered as a third authority source but proved unobtainable across the access paths tested. Getty was substituted before the confirmatory protocol was fixed, and the scope claim was narrowed accordingly.

A frame of 5,247 unique entities was frozen after capped enumeration of three property pairs. Each pair was paged with LIMIT/OFFSET without ORDER BY and capped at 3,000 returned rows. The resulting frame is therefore an engine-ordered slice rather than a probability sample of the full property-pair populations. From this frame, a 3,000-entity stratified analysis pool was formed. Because the pool size equalled the pre-specified target, the final selection stage retained all 3,000 entities. The realised design is consequently a post-stratified analysis of a frozen analysis pool rather than stratified probability sampling.

Entity and revision snapshots were acquired once and then treated as immutable. All reported analyses run from those archived snapshots. This design supports exact re-analysis of the frozen data but does not identify population-wide generalisation beyond the acquired slice.

3.2 Concordance definition and unit of analysis

A link is defined as concordant when Getty asserts skos:exactMatch between a Getty concept and the corresponding Library of Congress record. This is an external concordance signal supplied by Getty; it is not, by itself, evidence that the two authority assertions were independently established.

Each concordant pair contributes two classifiable statement-instances, one for each side of the pair. The final analytic set contains 773 concordant pairs and 1,546 statement-instances.

4 Method

4.1 Provenance signals and statement classification

Five descriptive provenance signals are extracted for each statement-instance: stated-in references (S-a), imported-from references (S-b), mediating identifiers within reference blocks (S-c), editor type (S-d), and a temporal variable (S-e). The pre-specified five-step classification rule uses S-a, S-b, S-c, and S-e to

assign one of three states: PROPAGATED, INDEPENDENT, or UNTRACEABLE. Editor type (S-d) is not a classifier input; it is retained only for diagnostics and auxiliary sensitivity analyses.

The classification rule, substantive decision threshold $\tau = 0.50$, γ floor, random seed, and bootstrap procedure were fixed and hash-verified before confirmatory data collection and before study outcomes were computed. A one-entity feasibility probe preceded the lock solely to verify that the required provenance fields were observable; it was not used to estimate study rates or inspect the confirmatory outcome. The 0.50 threshold represents a pre-specified substantive majority criterion for whether propagation is a material source of agreement. It is not a null-hypothesis significance threshold.

4.2 Estimands and partial identification

Three quantities organise the analysis. The protocol metric γ, originally labelled “coverage,” is operationally the hybrid-weighted traceability proportion within strata that contain analysed statements. It does not describe overall frame coverage because strata without an analysed statement contribute no rate estimate. The quantity θ is the post-stratified independence rate among traceable statements. The decision metric θ^{+} assigns every UNTRACEABLE statement to INDEPENDENT, a deliberately independence-favouring allocation that biases against the propagation hypothesis.

The unresolved class also defines the identification problem. Let p denote the propagated fraction of the untraceable remainder. The observed provenance does not identify p. The lower endpoint of the independence rate assigns all untraceable cases to PROPAGATED; the upper endpoint assigns all of them to INDEPENDENT. The resulting interval is therefore assumption-free with respect to the latent derivation of the untraceable statements. Any point estimate inside that interval would require an additional assumption about p or its distribution across strata.

4.3 Hybrid post-stratification and uncertainty

The weighting scheme is hybrid post-stratification. Entity-type mass is anchored to stage-1 property-pair population counts and within-pair type composition. By contrast, link_band and the pre-specified period dimension are estimated from the frozen 3,000-entity analysis pool because these quantities were unavailable for unacquired frame members. Reported rates are therefore standardised rates over the estimable slice, not simple prevalence estimates for all concordant links in the source populations.

Strata with no analysed statement remain unestimated. Their weight is excluded from the rate calculation, the remaining weights are renormalised, and the excluded mass is reported explicitly. This distinction is important because a stratum being represented in the analysis pool does not imply that it contributes an estimable concordant-statement rate.

Uncertainty is assessed with a percentile cluster bootstrap using Wikidata entities as the resampling unit, because the two statement-instances belonging to a concordant pair are not independent observations. The pre-specified bootstrap resamples entities globally. Because the point estimator is post-stratified, a stratum-preserving entity bootstrap is also reported as a robustness analysis.

4.4 Design safeguards and corrective diagnostics

Concordance detection is separated from provenance classification: population membership is determined without reading the provenance signals subsequently used for classification. This separation is enforced in the analysis implementation. Publication-facing release artifacts are also purged of person attributes and editor identities.

A subsequent implementation review identified one discrepancy in S-e. The field named first_asserted contained the entity's earliest revision timestamp rather than the target statement's first-appearance timestamp. The confirmatory baseline is retained for transparency, but no substantive interpretation is attached to S-e as a statement-level temporal-order measure. Two corrective analyses quantify the effect of this discrepancy: the six classifications decided by S-e are conservatively reassigned to UNTRACEABLE, and the affected period dimension is removed from the weighting scheme. Because these checks were defined after the discrepancy was discovered, they are reported as post hoc robustness analyses rather than confirmatory tests.

Additional estimand definitions, sensitivity details, influence diagnostics, and reproducibility information are provided in Online Resource 2.

4.5 Use of generative AI during manuscript preparation

Generative AI tools were used during manuscript preparation for literature exploration, manuscript organization and restructuring, language refinement, and reviewer-style critique. The author independently checked the cited sources, archived data, analysis code, numerical results, methodological decisions, and interpretations and takes full responsibility for the final manuscript. Generative AI was not treated as an author and was not used to create the primary study data or to replace execution of the reported analyses.

5 Results

5.1 Provenance traceability and observed derivation classes

Analysed statements occur in 11 of 27 post-stratification strata. Those 11 strata carry 41.6% of the total hybrid post-stratification weight; the remaining 58.4% has no concordant-statement rate estimate. Within the estimable strata, $\gamma = 0.607$, meaning that approximately 60.7% of hybrid-weighted observations are traceable under the classification rule. This quantity should not be read as 60.7% coverage of the full frame.

Across the 1,546 statement-instances, 728 are classified as PROPAGATED, 256 as INDEPENDENT, and 562 as UNTRACEABLE. Among the 984 traceable statements, $\theta = 0.289$ (95% bootstrap interval [0.255, 0.325]). Equivalently, approximately 71% of traceable agreement is propagated rather than independently

established. This rate describes the traceable subgroup of the estimable ULAN–LC-name slice; it does not describe the unresolved statements or cultural-heritage authority links generally. Figure 1 shows the observed class composition.

5.2 The independence rate is only partially identified

The 562 untraceable statements prevent a unique estimate of the overall independence rate. If every untraceable statement is assigned to INDEPENDENT, the upper endpoint is $\theta^{+}=0.556$. Assigning all untraceable statements to PROPAGATED gives the lower endpoint 0.163. The resulting identified interval is therefore:

[0.163, 0.556]

The interval spans the pre-specified substantive threshold $\tau = 0.50$. The principal result is thus not that independence lies above or below one-half, but that the available provenance does not identify which side of that threshold the underlying post-stratified independence rate occupies. Figure 2 visualises the identified set relative to the threshold and the traceable-only estimate.

Identification uncertainty and sampling uncertainty are distinct. Under the pre-specified entity-cluster bootstrap, the lower-endpoint estimator has a 95% percentile interval of approximately [0.145, 0.182], whereas the upper-endpoint estimator has an interval of [0.530, 0.578]. The combined range [0.145, 0.578] is descriptive only and is not presented as a formal confidence set for a partially identified parameter.

5.3 The threshold decision depends on a small unresolved propagated share

The pre-specified decision metric θ^{+} treats every UNTRACEABLE statement as INDEPENDENT, deliberately favouring independence. Under this allocation, $\theta^{+}=0.556$ (95% bootstrap interval [0.530, 0.578]), so the pre-specified majority rule provides no support for the propagation hypothesis. This outcome should not be interpreted as positive evidence that the links were independently curated.

The threshold result depends on the unidentified composition of the untraceable class. Across the tested allocations, $\tau = 0.50$ is reached when only about 7.7–16.9% of the untraceable statements are propagated; the exact break-even point depends on where those cases occur across strata. Table 1 summarises the corresponding sensitivity profile.

Table 1
Sensitivity of the post-stratified independence rate to the propagated fraction of untraceable statements.

Propagated fraction among untraceable cases (p)	Independence rate	Relation to $\tau = 0.50$
0.00	0.556	above
0.08–0.17	≈ 0.50	break-even
0.20	0.481	below
0.71	0.278	below
1.00	0.163	below

Using the propagated share observed among traceable statements as an estimate of p would require an untestable exchangeability assumption. Traceable and untraceable statements differ on auxiliary variables not used by the classifier: total variation distance is 0.245 for editor type and 0.207 for statement side/source, while divergence by link-count band is smaller at 0.062. Moreover, 547 of the 562 untraceable statements have no reference at all. MAR-type extrapolations based on valid auxiliary covariates imply propagated fractions of approximately 0.705–0.788, but these are sensitivity scenarios rather than estimates because the missingness mechanism is not identified.

5.4 Robustness and corrective analyses

The identification conclusion is stable across the corrective and alternative analyses examined here, although the upper endpoint itself is not invariant to modelling choices. Table 2 collects the main robustness results.

Table 2
Robustness and corrective analyses for the identified independence interval.

Analysis	Lower endpoint	Upper endpoint / θ^+	95% interval for upper endpoint	Relation to $\tau = 0.50$
Pre-specified hybrid post-stratification	0.163	0.556	[0.530, 0.578]	above
S-e-decided cases → UNTRACEABLE	0.158	0.556	unchanged	above
period dimension removed	0.167	0.543	[0.517, 0.569]	above
S-e correction + period removed	0.162	0.543	[0.517, 0.569]	above
Shifted link-count bands	—	0.545	—	above
Unweighted observed statements	0.166	0.529	[0.503, 0.556]	above

Only six INDEPENDENT labels were determined by S-e. Reassigning those six cases to UNTRACEABLE lowers the identified-set lower endpoint from 0.163 to 0.158 while leaving the upper endpoint unchanged. Removing the affected period dimension produces an upper endpoint of 0.543, still above 0.50.

The alternative stratum-preserving entity bootstrap yields $\theta = 0.290$ with a 95% interval of [0.257, 0.324] and $\theta^{+}=0.556$ with an interval of [0.530, 0.581], closely matching the pre-specified global entity bootstrap. The resampling design is therefore not driving the identification result.

The margin above 0.50 is nevertheless concentrated. The two largest estimable strata carry 80.2% of the renormalised hybrid weight. Omitting the stratum person|4–6|2015–2019, which carries 26.2% of that weight, reduces θ^{+} to 0.496. The appropriate robustness claim is therefore not that the threshold result is diffuse across strata, but that the broader partial-identification conclusion survives the corrective analyses considered here.

5.5 The empirical mechanism is concentrated in ULAN-linked persons

A separate census of all 3,081 Getty identifiers in the analysis pool clarifies the scope of the concordance mechanism (Table 3).

Table 3
Getty identifier census and observed Library of Congress targets in the analysis pool.

Vocabulary	Identifiers	exactMatch triples	To LC names	To LCSH subjects
ULAN	1,092	1,325	782	0
AAT	948	1,030	0	49
TGN	1,041	0	0	0

Under the study's LC-name definition, Getty-to-LC-name concordance occurs essentially only through ULAN. AAT contains 49 id.loc.gov subject-heading targets rather than LC name-authority targets. For 48 of those, the target coincides with the entity's P244 value. AAT can therefore corroborate an LC authority in a broader subject-heading sense even though it contributes no records under the names-only definition. TGN contributes no skos:exactMatch triples in the census.

Consistent with this mechanism, 1,540 of the 1,546 analysed statements concern persons. The numerical rate claims in this study therefore apply to the estimable ULAN–LC-name slice, dominated by persons, rather than to Getty vocabularies or cultural-heritage authority links as a whole.

6 Discussion

6.1 Agreement and corroboration answer different evidential questions

The results separate two concepts that are easy to conflate in linked-data practice. Concordance establishes that two authority pathways agree on an identifier. Independent corroboration requires the stronger condition that the agreeing assertions were not merely transmitted through the same provenance chain. In this study, provenance directly contradicts independence for a substantial share of the traceable statements: approximately 71% are classified as propagated. For the remaining statements whose derivation cannot be resolved, the evidence supports neither an independence label nor a propagation label.

This distinction does not reduce the operational value of reconciliation. A copied identifier can still improve interoperability, discovery, deduplication, and record linkage. The evidential consequence is narrower: repeated agreement should not automatically be counted as repeated independent support. Reconciliation quality and corroborative independence are therefore complementary but different properties.

6.2 Missing provenance changes what can be identified

The 36.4% untraceable share is not treated as ordinary missing metadata that can be innocuously filled in. Its latent composition determines whether the post-stratified independence rate falls above or below the pre-specified majority threshold, and the threshold can reverse when only a modest share of the unresolved cases is propagated. Differences between traceable and untraceable statements further weaken the case for assuming that the two groups are exchangeable.

Partial identification makes this uncertainty explicit. Rather than selecting one missingness model and presenting the resulting point estimate as if it were observed, the identified interval states what the provenance supports without an additional assumption about unresolved derivation. Better provenance would therefore do more than improve documentation: it would narrow the identified set and increase the range of scientific claims the data can support.

6.3 Implications for digital-library, authority-control, and knowledge-graph workflows

The findings suggest a provenance-aware interpretation of cross-authority agreement in digital-library metadata infrastructures. Workflows that use agreement across authority systems as an evidential signal should, where possible, distinguish at least three states: independently established concordance, traceably propagated concordance, and concordance whose derivation remains unresolved. A confidence model that ignores these distinctions risks assigning the same evidential weight to independent confirmation and to repeated information flowing through an import chain.

For digital-library and authority-data publishers, preserving import provenance, reference sources, and statement-introduction history can therefore improve both auditability and inferential value. For downstream services that aggregate authority links for discovery, entity resolution, metadata enrichment, or trust assessment, provenance-aware aggregation can prevent dependence from being mistaken for replication. These implications extend beyond Wikidata, but the numerical rates reported here should not: they belong to the specific estimable ULAN–LC-name slice analysed in this study.

6.4 Scope of the contribution

The study does not show that most cultural-heritage authority links are propagated, that Wikidata is uniquely affected by source dependence, or that the unresolved statements are likely to belong to one latent class. The frame construction also prevents population-wide inference to all relevant Wikidata authority links.

The narrower conclusion is more defensible and more reusable. In a real cross-authority setting, value agreement coexists with substantial traceable propagation and substantial unresolved derivation. Under those conditions, independence becomes a partial-identification problem. The methodological lesson can transfer to other linked-data environments even when the rates themselves do not.

7 Limitations

The principal limitation is unresolved derivation. The propagated fraction of the UNTRACEABLE class, p, is not identified, so the study cannot recover a single overall independence rate without additional assumptions or evidence. A revision-history extension was developed after the confirmatory analysis to investigate these cases more deeply, but manual adjudication from that extension is not part of the results reported here. The primary analysis therefore retains the assumption-free identified interval.

External validity is limited by both frame construction and the realised design. Frame enumeration used capped LIMIT/OFFSET paging without ORDER BY, making the frozen frame an engine-ordered slice rather than a probability sample of the full property-pair populations. Because the 3,000-entity stratified pool equalled the pre-specified target size, the final selection stage retained the entire pool. The realised design is therefore post-stratified analysis rather than stratified probability sampling, and generalisation beyond the frozen slice is not identified.

Analytic support narrows further at concordance. Sixteen of 27 strata contain no analysed statement, so all rate estimates rest on 41.6% of the total hybrid post-stratification weight; 58.4% carries no rate estimate. The γ metric is conditional on estimable strata and should not be interpreted as overall frame coverage. Places and concepts account for most of the missing weight, while the resulting analytic set is overwhelmingly composed of persons in ULAN concordant with LC name authorities.

The temporal signal requires specific qualification. S-e was populated with the entity's earliest revision timestamp rather than the target statement's first-appearance timestamp and therefore cannot support statement-level temporal-order claims. Only six INDEPENDENT classifications were determined by S-e.

Reassigning those cases to UNTRACEABLE leaves the upper identified endpoint unchanged at 0.556, and removing the affected period dimension yields an upper endpoint of 0.543. The discrepancy therefore changes interpretation and some sensitivity values but not the conclusion that the substantive threshold lies inside the identified set.

The post-stratification weights are partly estimated. Entity-type mass is anchored to property-pair population counts, whereas link-count-band and period composition are derived from the frozen analysis pool. The target is consequently a standardised rate over the estimable slice rather than raw prevalence among all concordant records. The upper endpoint is 0.529 in the unweighted analysis and 0.543 when the period dimension is removed, showing that the margin above 0.50 is sensitive to weighting choices even though the identification conclusion is preserved.

Influence is also concentrated. Removing one of the 11 estimable strata, person|4–6|2015–2019, is sufficient to reduce θ^{+} below 0.50. This does not invalidate the estimator, but it prevents interpreting the pre-specified threshold result as broadly distributed across strata.

Finally, the design can detect only propagation represented in the available provenance signals. Second-order propagation—where an apparently corroborating source is itself derivative—remains invisible and therefore tends to under-detect propagation. The 150/150 validation re-derivations are an internal self-consistency check rather than inter-rater reliability, and reproducibility is limited to re-analysis from archived snapshots because the underlying sources remain live and may change.

8 Conclusion

This study asked a narrow evidential question: when authority identifiers agree, how much of that agreement can be treated as independently established? In the analysed Wikidata-mediated ULAN–LC-name slice, the available provenance does not support a single overall answer. The identified independence interval is [0.163, 0.556], spanning the substantive threshold of 0.50, and corrective analyses for the temporal-variable discrepancy, alternative weighting, and bootstrap design do not remove that ambiguity.

The clearest direct finding is that traceable agreement is more often propagated than independently established: approximately 71% of traceable statements are classified as propagated. At the same time, 36.4% of analysed statements remain untraceable, and only a modest propagated share of that unresolved remainder is needed to move the threshold result. The evidence therefore supports neither a general claim of independent curation nor a general claim that all concordance is derivative.

The practical implication is straightforward. Agreement can be valuable for reconciliation without constituting independent corroboration. When agreement is used as evidence, provenance and source dependence should be modelled explicitly. Missing provenance is not only a documentation defect; in settings such as this one, it determines what can and cannot be identified from the data.

Declarations

Competing Interests

The author has no relevant financial or non-financial interests to disclose.

Ethics Approval and Consent to Participate

Not applicable. The study analyses publicly available authority and knowledge-graph metadata and did not involve interaction with human participants or interventions.

Use of AI-assisted tools

Generative AI tools were used during manuscript preparation for literature exploration, manuscript organization and restructuring, language refinement, and reviewer-style critique. The author independently checked the cited sources, archived data, analysis code, numerical results, methodological decisions, and interpretations and takes full responsibility for the final manuscript. Generative AI was not treated as an author and was not used to create the primary study data or to replace execution of the reported analyses.

Funding

No funding was received for conducting this study.

Author Contributions

The sole author was responsible for study conception, methodology, data curation, formal analysis, interpretation, and preparation and revision of the manuscript, and approved the final version.

Data and Code Availability

All reported estimates derive from archived, dated snapshots. A reproducibility package containing the analysis code used for the reported estimates, derived evidence files, analysis specifications, figure-generation code, and checksums is supplied with the submission as Online Resource 1. The corresponding GitHub repository is private during peer review and is planned for public release upon publication. Re-analysis from the archived evidence reproduces the reported calculations; reacquisition from the live sources is not expected to reproduce an identical snapshot. Supplementary methodological and robustness details are supplied as Online Resource 2.

References

1. Tharani K (2021) Much more than a mere technology: A systematic review of wikidata in libraries. J Acad Librariansh 47(2):102326. https://doi.org/10.1016/j.acalib.2021.102326
2. Bianchini C, Bargioni S, Pellizzari CC, Beyond VIAF (2021) Wikidata as a complementary tool for authority control in libraries. Information Technology and Libraries. ;40(2). https://doi.org/10.6017/ital.v40i2.12959
3. Zhao F (2023) A systematic review of wikidata in digital humanities projects. Digit Scholarsh Humanit 38(2):852–874. https://doi.org/10.1093/llc/fqac083
4. Candela G, Cuper M, Holownia O, Gabriels N, Dobreva M, Mahey M (2024) A systematic review of wikidata in GLAM institutions: A labs approach. In: Linking theory and practice of digital libraries: 28th international conference on theory and practice of digital libraries, TPDL 2024. Springer; pp. 34–50. https://doi.org/10.1007/978-3-031-72440-4_4
5. Pellizzari di San Girolamo CC (2024) The reconciliation of SBN authority records with wikidata. Progresses and perspectives after a decade of work (2013–2023). JLISit 15(1):33–44. https://doi.org/10.36253/jlis.it-573
6. Foster A, Norling S, Westerby MJ (2026) Towards a linked open index: Reconciling museum records to wikidata for index of american design constituents. J Open Humanit Data 12:1. https://doi.org/10.5334/johd.426
7. Trognitz M, Mandell RA, Štuhec S, Palacz J (2026) Wikidata as a reconciliation anchor: Curating data for long-term preservation. J Open Humanit Data 12:27. https://doi.org/10.5334/johd.450
8. Pellizzari di San Girolamo CC (2026) Wikidata and authority files: Reconciliation and cooperation procedures. JLISit 17(2):33–50. https://doi.org/10.36253/jlis.it-698
9. Dong XL, Berti-Equille L, Srivastava D (2009) Integrating conflicting data: The role of source dependence. Proceedings of the VLDB Endowment. ;2(1):550–61. https://doi.org/10.14778/1687627.1687690
10. Li Y, Gao J, Meng C, Li Q, Su L, Zhao B et al (2016) A survey on truth discovery. ACM SIGKDD Explorations Newsl 17(2):1–16. https://doi.org/10.1145/2897350.2897352
11. Zhang H, Wang S, Chen L, Li X, Gao Q, Sheng QZ (2026) Reliable truth discovery for dynamic and dependent sources. IEEE Trans Knowl Data Eng 38(1):546–558. https://doi.org/10.1109/TKDE.2025.3631376
12. Daquino M, Mambelli F, Kozlov A (2026) Many-to-many. Usability challenges of entity reconciliation in art history and photographic studies. J Doc 82(2):362–380. https://doi.org/10.1108/JD-09-2025-0284
13. Zhang J, Yan X (2026) Multilingual crosswalks for chinese intangible cultural heritage: Mapping china's UNESCO-inscribed elements across wikidata, LCSH, and getty AAT. Knowl Organ 53(4):52928. https://doi.org/10.31083/KO52928

14. McKenna L, Debruyne C, O'Sullivan D (2022) Using linked data to create provenance-rich metadata interlinks: The design and evaluation of the NAISC-l interlinking framework for libraries, archives and museums. AI Soc 37:921–947. https://doi.org/10.1007/s00146-021-01373-z
15. Hosseini Beghaeiraveri SA, Gray A, McNeill F (2024) RQSS: Referencing quality scoring system for wikidata. Semantic Web 15(6):2419–2475. https://doi.org/10.3233/SW-243695
16. Di Pasquale A, Pasqual V, Tomasi F, Vitali F (2024) On assessing weaker logical status claims in wikidata cultural heritage records. Semantic Web 15(6):2395–2417. https://doi.org/10.3233/SW-243686
17. Menotti L, Marchesin S, Giachelle F, Silvello G (2025) Provenance-driven nanopublications: Representing source lineage and trust networks for multi-source assertions. Int J Digit Libr 26:24. https://doi.org/10.1007/s00799-025-00431-x
18. Shenoy K, Ilievski F, Garijo D, Schwabe D, Szekely P (2022) A study of the quality of wikidata. J Web Semant 72:100679. https://doi.org/10.1016/j.websem.2021.100679

Figures

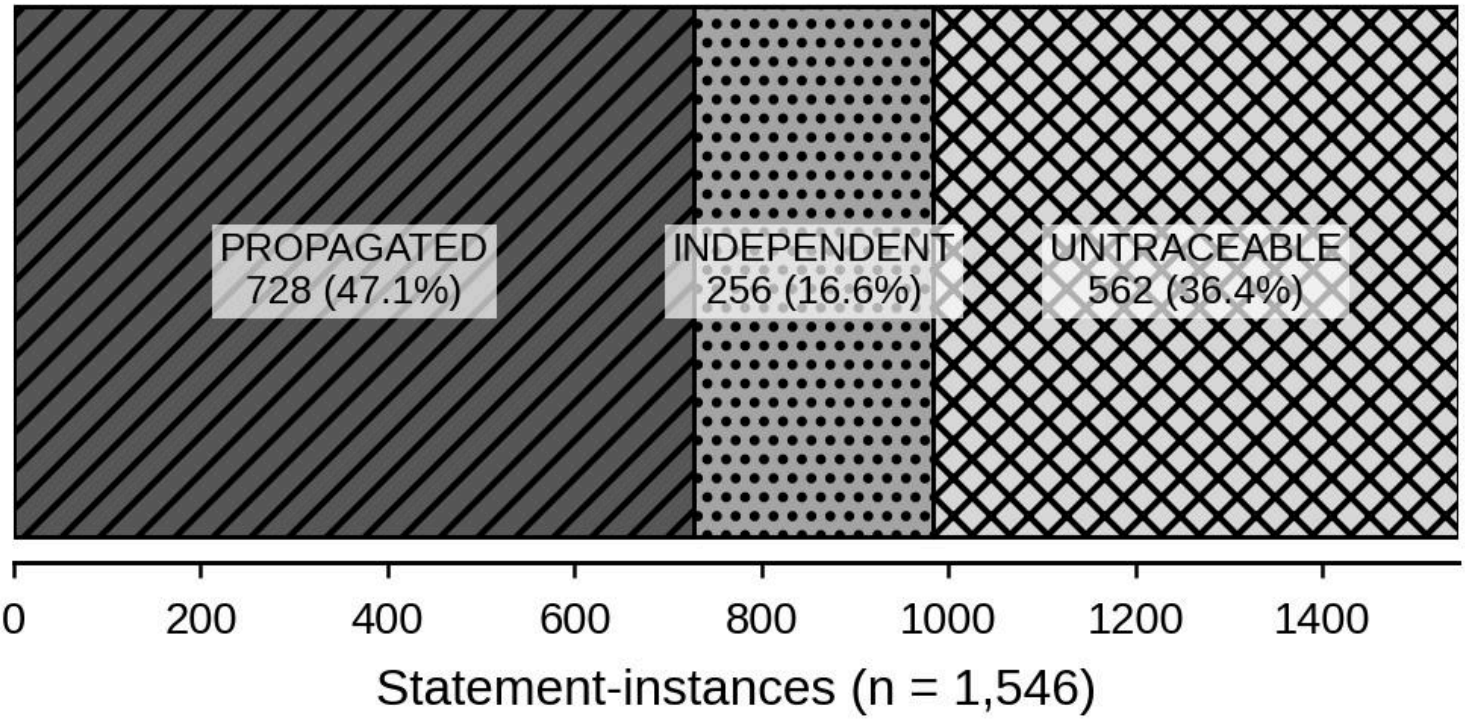


Figure 1

Observed provenance-class composition across the 1,546 analysed statement-instances. Traceable statements comprise the PROPAGATED and INDEPENDENTclasses; UNTRACEABLE statements are retained as unresolved rather than forced into either class

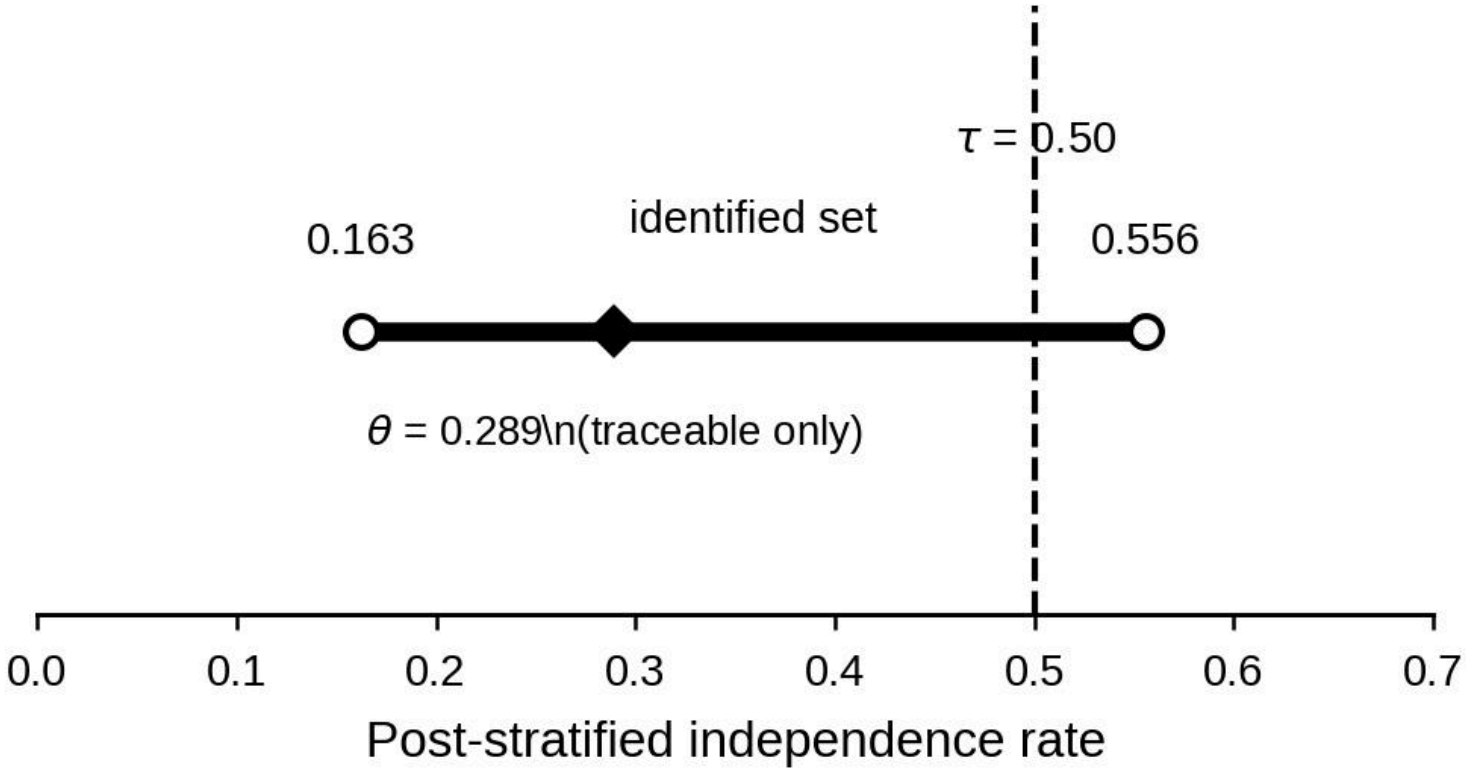


Figure 2

Partial identification of the post-stratified independence rate. The identified interval [0.163, 0.556] spans the substantive threshold τ=0.50; the traceable-only estimate θ=0.289 is shown for context and is not an overall point estimate

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialIJDL.pdf