Supplementary Information

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

International Journal on Digital Libraries

Emin Talip Demirkiran
Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, 26555 Eskisehir, Türkiye
Corresponding author: etd@eskisehir.edu.tr
ORCID: https://orcid.org/0000-0001-6063-8178

S1. Scope and analysis unit

The supplementary material documents the definitions, robustness checks, and reproducibility details supporting the main manuscript. The empirical rate claims concern the estimable Wikidata-mediated Getty ULAN–Library of Congress name-authority concordant slice and are dominated by persons. They are not population estimates for all cultural-heritage authority links.

The unit of analysis is the statement-instance. The analysis contains 1,546 statement-instances associated with 773 concordant entity pairs from a frozen 3,000-entity analysis pool. Provenance classes are PROPAGATED, INDEPENDENT, and UNTRACEABLE. Missing derivation evidence remains unresolved rather than being forced into either of the first two categories.

S2. Estimands and identification

Table S1. Core estimands and interpretation

Quantity	Definition	Interpretation
γ	Hybrid-weighted traceable share within estimable strata	Descriptive traceability quantity; not full-frame coverage
θ	Independence rate among traceable statements	Conditional on statements that can be classified
Lower endpoint	UNTRACEABLE assigned to PROPAGATED	Most propagation-favouring endpoint: 0.163
θ+ / upper endpoint	UNTRACEABLE assigned to INDEPENDENT	Most independence-favouring endpoint: 0.556
p	Propagated fraction within UNTRACEABLE	Not identified by the primary design
τ	Pre-specified substantive threshold	0.50; a majority threshold, not a significance level

Because p is not identified, the overall post-stratified independence rate is partially identified. The assumption-free identified interval is [0.163, 0.556]. Sampling uncertainty for the endpoint estimators is reported separately from identification uncertainty.

S3. Observed provenance classes and traceability

Table S2. Statement-level provenance classes

Class	Count	Share of all statements
PROPAGATED	728	47.1%
INDEPENDENT	256	16.6%
UNTRACEABLE	562	36.4%

Among the 984 traceable statements, the weighted independence estimate is θ=0.289 (95% bootstrap interval [0.255, 0.325]); equivalently, approximately 71% of traceable agreement is propagated. Of the 562 untraceable statements, 547 have no reference.

S4. Sensitivity and robustness analyses

Table S3. Robustness summary

Analysis	Lower endpoint	Upper endpoint	Upper 95% interval	Threshold relation
Pre-specified hybrid post-stratification	0.163	0.556	[0.530, 0.578]	spans threshold in identified set
S-e-decided cases reclassified as UNTRACEABLE	0.158	0.556	unchanged	spans threshold
period dimension removed	0.167	0.543	[0.517, 0.569]	spans threshold
S-e correction + period removed	0.162	0.543	[0.517, 0.569]	spans threshold

Analysis	Lower endpoint	Upper endpoint	Upper 95% interval	Threshold relation
Shifted link-count bands	—	0.545	—	spans threshold
Unweighted observed statements	0.166	0.529	[0.503, 0.556]	spans threshold

The pre-specified majority threshold is reached when only about 7.7–16.9% of the untraceable statements are propagated, depending on their allocation across strata. A single pooled break-even value is therefore not treated as uniquely identified. MAR-type extrapolations using auxiliary variables yield propagated shares of approximately 0.705–0.788; these are scenarios rather than estimates because the missingness mechanism is not identified.

S5. Weight support and influence diagnostics

Analysed statements occur in 11 of 27 post-stratification strata. Those estimable strata carry 41.6% of the total hybrid post-stratification weight; 58.4% carries no concordant-statement rate estimate. The weighting system is hybrid: entity-type mass is anchored to property-pair population counts, whereas link-count-band and period composition are derived from the frozen analysis pool.

Table S4. Key influence diagnostic

Diagnostic	Result
Two largest estimable strata	80.2% of renormalised hybrid weight
Drop person|4-6|2015-2019	upper endpoint decreases to 0.496
Unweighted upper endpoint	0.529
Period-collapsed upper endpoint	0.543

The threshold margin is therefore concentrated rather than diffuse across strata. This limitation does not change the broader partial-identification result, because both sides of the 0.50 threshold remain compatible with the observed provenance under the identified set.

S6. Temporal-variable corrective analysis

The temporal signal S-e was populated with the entity's earliest revision timestamp rather than the target statement's first-appearance timestamp. It must therefore be interpreted as an entity-inception proxy rather than a statement-level assertion date. Six INDEPENDENT labels were determined by S-e. Reassigning those cases to UNTRACEABLE lowers the lower endpoint from 0.163 to 0.158 and leaves the upper endpoint at 0.556. Removing the affected period dimension gives an upper endpoint of 0.543. The discrepancy changes interpretation and some sensitivity values but does not alter the conclusion that the threshold lies inside the identified set.

S7. Concordance scope census

Table S5. Getty identifier census 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 names-only definition used in the study, Getty-to-LC-name concordance occurs essentially only through ULAN. AAT contributes 49 Library of Congress subject-heading targets, 48 of which coincide with the entity's P244 value, but these are outside the primary names-only definition. TGN contributes no skos:exactMatch triples in the census. Of the 1,546 analysed statement-instances, 1,540 concern persons.

S8. Reproducibility and archived evidence

Online Resource 1 contains a self-contained reproducibility package with cleaned analysis scripts, derived statement and stratum evidence, analysis specifications, figure-generation code, expected results, and SHA-256 checksums. Re-running the analysis on these archived evidence files reproduces the reported class counts and

primary endpoint estimates. Live-source reacquisition is not expected to recreate an identical snapshot because the underlying services are mutable.

Table S6. Main reproducibility artefacts in Online Resource 1

Path	Purpose
code/reproduce_primary.py	Recomputes class counts, θ, lower and upper endpoints, and bootstrap intervals
code/generate_figures.py	Regenerates manuscript figures from reported quantities
data/statements_release.jsonl	Statement-level derived provenance classes
data/strata.jsonl	Frozen stratum assignment for each entity
data/weight_provenance.json	Hybrid post-stratification weights and provenance
data/sensitivity_prespecified.json	Archived sensitivity-analysis outputs
protocol/analysis_specification.json	Publication-facing analysis specification and caveats
SHA256SUMS.txt	Integrity checksums for repository files

S9. Data governance

The reproducibility package contains derived identifiers and provenance classifications needed for analysis; it does not include raw person-attribute payloads. The study uses public authority and knowledge-graph resources. No authentication bypass or access-control circumvention was used. Repository release is planned after publication, while the review package supplied with the submission provides the material needed to inspect the reported analysis.