Who Gets Linked? Protocol-Conditioned Linkability and Trace-Level Evidence in Italian Cultural-Heritage Entity Linking

Alfredo Gioacchino MariaPio Vecchio, Andrea Pietro Arena*, Sofia Marilina Glorioso, Nicole Dalia Cilia and Salvatore Sorce

University of Enna "Kore", Department of Engineering and Architecture, 94100 Enna, Italy

Abstract

Entity linking for cultural-heritage records is usually evaluated by whether a system selects the intended authority target. This preliminary study asks an earlier question: under a bounded search protocol, for which records is a target established at all? On 450 ArCo person and cultural-property records, we separate target establishment, candidate generation, link selection, and trace-level evidence, distinguishing three stage-specific conditions: no reference target established, an established target absent from the candidate set, and scoring-stage abstention. A single machine annotator produced a provisional reference: target-establishment rates ranged from 0 to 0.667 across five record profiles, while the automated harvest returned no candidate for 58% of records, so a system NIL may reflect retrieval failure rather than verified knowledge-base absence. A lexical scorer with a temporal veto and a bi-encoder without one both emitted more wrong links than correct ones, with no conclusive F1 difference, and both were dominated by a trivial always-abstain baseline under an asymmetric cost. The apparent completeness of warning coverage was driven by one systematically unavailable evidence field. Human validation was not included; a targeted audit is the immediate next step, required before these rates can be read beyond this protocol and annotator.

Keywords

entity linking, cultural heritage, authority files, linkability, NIL detection, evidence availability

1. Introduction

Linking archival and museum records to authority infrastructures such as Wikidata is usually framed as an accuracy problem: how often does a system select the right target? This study starts one step earlier. For many records, a bounded search establishes no target at all, and different record profiles are resolved at different rates. For an archive agent recorded only as *Phot. Klopfenstein*, leaving the record unlinked or assigning it to a famous homonym depends on engineering defaults that are rarely measured directly. In cultural-heritage infrastructures, these choices also distribute visibility and attribution across the digital record through which cultural memory is accessed.

We call the share of resolved records for which a bounded procedure establishes a target *protocol-conditioned linkability* (Section 3). This staged view motivates three questions. First, how often does *the reference annotation protocol*—not a linking system—establish a target at all across the observed record profiles, and how sensitive are those rates to the protocol's own abstention? This question is prior to, and independent of, whether any system subsequently links or abstains. Second, how often does the shared candidate harvest retrieve the targets established by the reference protocol, and how accurately do the two linker configurations select a candidate or abstain on the decidable records, under strict link-level scoring and under an illustrative asymmetric cost that penalises wrong links more heavily? Third, when

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, workshop of AIxIA 2026, October 6–9, 2026, Perugia, Italy

*Corresponding author.

alfredogioacchinomariapio.vecchio@unikorestudent.it (A. G. M. Vecchio); andreapietro.arena@unikorestudent.it (A. P. Arena); sofiamarilina.glorioso@unikorestudent.it (S. M. Glorioso); nicoledalia.cilia@unikore.it (N. D. Cilia); salvatore.sorce@unikore.it (S. Sorce)

0009-0000-6018-2756 (A. G. M. Vecchio); 0009-0001-6138-615X (A. P. Arena); 0009-0000-8371-5543 (S. M. Glorioso); 0000-0002-9631-5802 (N. D. Cilia); 0000-0003-1976-031X (S. Sorce)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4314/ReDH7.pdf

a system emits an incorrect link, which forms of potentially disconfirming evidence—counter-evidence, unavailable corroboration, or ambiguity—are actually exposed by its trace? The first question provides a task-level measurement of coverage differences documented qualitatively in authority-infrastructure scholarship [1, 2]; the second separates candidate availability from link-selection behaviour; and the third connects system evaluation to questions of provenance and contestation in digital humanities [3, 4].

The contribution is a staged evaluation that distinguishes three stage-specific conditions that are often conflated when interpreting NIL and linking failures: **N1**, the annotation protocol establishes no reference target; **N2**, an established reference target is absent from the harvested candidate set; and **N3**, the scorer receives candidates but emits NIL. N1 is a property of the reference procedure, N2 of retrieval, and N3 of link selection. These conditions are reported separately rather than treated as interchangeable forms of knowledge-base absence.

2. Related work

Linkage error varies across groups in voter registries, hospital databases, and fair entity-resolution studies [5, 6, 7]. Our quantity precedes scoring error: whether a bounded protocol establishes a target at all. Explainable entity resolution interprets matching decisions [8], whereas our narrower audit records which potentially disconfirming signals a trace makes available. NIL-aware evaluation studies out-of-knowledge-base discovery [9]; here reference NILs are reported with candidate availability rather than as a residual category.

Digital-humanities research documents authority-file coverage, gender skews, classificatory harm, and the difficulty of representing historical lives in Wikidata [1, 2, 3, 4]. Historical entity-linking work largely reports aggregate performance [10]; archival disambiguation [11] and museum authority control [12] are closest to our setting.

3. Corpus, strata, and reference set

From the ArCo SPARQL endpoint, deterministic MD5-seeded ordering (`redh2026-d1-seed42`) drew 800 agents and 800 works, yielding 1,586 unique records. Four source-side dimensions were pre-specified: entity type, richness (at most 2 populated fields: poor; 3–5: medium; at least 6: rich), dating, and location. Gender was dropped because it was never populated. The dimensions produce five profiles and are confounded: every person is unlocated and not rich, and richness coincides with dating for both persons and works, so their effects cannot be estimated independently.

A proportional stratified sample of 450 records (`redh2026-d1-gold-seed7`) was annotated by an LLM agent recorded only under the internal pipeline label *Fable 5*. It applied a frozen codebook to per-record evidence dossiers using the label, up to three aliases, one structured query, a VIAF lookup while available, and a five-minute cap. The provider, model version, prompt, and decoding parameters were not recorded, so the pass is documented but not re-executable. The codebook was SHA-256 committed before either scorer was implemented, preventing retrospective fitting. It required UNDECIDABLE whenever the decision tree did not resolve; 101 records (22%) received that verdict.

No human validation was included in this preliminary phase, whose aim was to specify and execute the staged measurement before incurring the cost of expert adjudication. The resulting reference is a provisional output of one bounded protocol, not ground truth, and all reported rates remain conditional on it. The protocol's explicit UNDECIDABLE verdict is part of the measurement rather than missing data: the bounds in Table 1 show how far each profile rate could move under the two extreme assignments of those abstentions.

For persons, a target is the same individual. For cultural properties, it is a standalone Wikidata item denoting the same property at the ArCo record's granularity; items for a creator, attributed master, institution, place, or generic type do not qualify. Failure to establish such a target may reflect knowledge-base absence, search failure, or a granularity mismatch, which this study does not decompose. The

definition is deliberately strict: a plausible link to a related entity does not count as identifying the catalogued object.

The annotator's search and the automated candidate harvest are distinct. The harvest shared by both linkers used Wikidata prefix search with orthographic variants, full-text search for residual zero-candidate records, and enrichment with types, dates, and VIAF property P214. Seven of 24 reference targets were unreachable: five records had no candidate and two only wrong candidates. The VIAF AutoSuggest API was discontinued during harvesting, and a missing P214 claim is not evidence of absence. Overall, 262 records (58%) had no candidate. The harvest is deliberately a simple baseline, held constant so that both linkers see identical candidate sets; this 58% therefore characterises the retrieval stage of this pipeline, not the coverage of Wikidata.

4. Systems

S1 is a CPU-compatible lexical scorer, $0.7 \cdot$ token-set ratio $+ 0.3 \cdot$ character ratio on a 0–100 scale, with a $\pm$10-year tolerance for *circa* dates and a hard veto on disjoint ranges. S2 is the multilingual bi-encoder `paraphrase-multilingual-MiniLM-L12-v2` (revision `e8f8c21`), using cosine similarity between record label plus date literals and candidate label plus description, without temporal filtering. ArCo dating resource identifiers are not parsed as years. Differing in architecture, features, and temporal handling, the two are compared as complete configurations rather than as an ablation.

Thresholds were selected by F1 on a stratified development half and frozen for test: $\theta = 70$ for S1 over $\{60, 65, \ldots, 95\}$ and $\theta = 0.65$ for S2 over $\{0.50, 0.55, \ldots, 0.90\}$. Versions, seeds, and input hashes are recorded in the manifest.

A minimal trace reports the selected score and runner-up margin. A negative-evidence emitter additionally reports temporal status (compatible, conflicting, or not performable) and nearby competitors. It is an external audit layer and does not expose either scorer's internal reasoning. The frozen checklist marks R1 when available evidence directly contradicts the selected candidate, R2 when the temporal check cannot be performed, and R3 when a competitor falls within 5 S1 points or 0.05 S2 cosine. R2 denotes unavailable corroboration, not proof of error. Under strict link-level scoring, a wrong identifier on a record with a reference target counts as both FP and FN.

Table 1
Protocol-conditioned linkability by record profile, test-set linker performance, and trace warnings over each system's false positives. *und.*: undecidable; *retr.*: established targets retrieved by the harvest; *Rate*: link/$(n - \text{und.})$; bounds assign all undecidable records to either outcome. CIs are Wilson for rates and percentile bootstrap for F1.

Record profile	n	und.	link	retr.	Rate [CI_{95}]	Bounds
person, medium *(dated, unlocated)*	24	0	16	11/16	0.667 [0.47, 0.82]	[0.67, 0.67]
person, poor *(undated, unlocated)*	199	83	5	3/5	0.043 [0.02, 0.10]	[0.03, 0.44]
work, poor *(undated, unlocated)*	35	18	3	3/3	0.176 [0.06, 0.41]	[0.09, 0.60]
work, medium *(undated, located)*	47	0	0	–	0.000 [0.00, 0.08]	[0.00, 0.00]
work, rich *(dated, located)*	145	0	0	–	0.000 [0.00, 0.03]	[0.00, 0.00]
all records	450	101	24	17/24	0.069 [0.05, 0.10]	[0.05, 0.28]

	S1 lex.-temp.	S2 bi-encoder
TP / FP / FN	9 / 34 / 4	9 / 24 / 4
Precision / Recall	0.209 / 0.692	0.273 / 0.692
F1 [CI_{95}]	0.321 [0.17, 0.49]	0.391 [0.20, 0.57]
Trace warnings, count (rate) over each system's own 34 resp. 24 false positives		
R1 counter-evidence *(conflict on selected candidate)*	0/34 (0.000)	0/24 (0.000)
R2 insufficiency *(temporal check not performable)*	34/34 (1.000)	24/24 (1.000)
R3 ambiguity *(competing candidate in margin)*	19/34 (0.559)	8/24 (0.333)

5. Results

Of 450 records, 349 are decidable: 24 LINK and 325 NIL. That NIL dominates the decidable set (93%) is itself a measurement, not a shortcoming of the design: it quantifies how sparsely this ArCo sample is resolvable against Wikidata under a bounded search, and it is why the abstention constraint below cannot be satisfied. The intervals in Table 1 (1,000 bootstrap resamples for F1) quantify sampling uncertainty, not annotation error.

Candidate recall is $17/24 = 0.708$ [0.508, 0.851], so N2 affects 7 of 24 established targets. Of 325 reference NILs, 323 carry the legacy label `kb_absent`, interpreted here only as target not established (N1); 236 of them also had no harvested candidate. Reference NIL and retrieval failure are therefore distinct, and neither verifies knowledge-base absence.

Protocol-conditioned linkability ranges from 0.667 for 24 dated person records to 0 for both located-work profiles ($n = 47$ and $n = 145$). These profiles contain no undecidable records, but candidate generation is uneven: 16/47 medium works and 111/145 rich works had no candidate. Their zero rates may reflect knowledge-base absence, annotation-search failure, or granularity mismatch; candidate scarcity separately constrains linker reachability.

On test ($n = 177$, 13 positives), both systems recover 9 targets. S1 emits 34 false positives and S2 24, against 9 true positives each, so neither is suitable for deployment at its F1-selected point. Their F1 difference is $+0.070$ $[-0.061, 0.190]$, which is inconclusive with only 13 positives. Profile macro-F1 is omitted because profiles without positives require an arbitrary zero-denominator convention and contain most false positives. Those errors are informative in their own right. Under strict scoring each of the 4 false negatives is either an abstention or a wrong target on a positive record, the latter also counting as a false positive; hence at least 30 of S1's 34 and 20 of S2's 24 false positives fall on reference-NIL records, a false-link rate over the 164 non-positive test records of at least 0.183 and 0.122. The systems are not merely imprecise on the 13 positives: they convert records the protocol left unresolved into asserted identifications.

The pre-specified diagnostic $C = 5\,FP + FN$ makes wrong links more costly; under strict scoring, a wrong target on a positive record contributes to both terms and costs 6. Its planned 60% abstention cap was incompatible with a decidable set containing 93% reference NILs: no grid point met it. Unconstrained minima reduced C from 174 to 97 for S1 and from 124 to 18 for S2, but required 89.8% and 99.4% abstention. A trivial always-abstain baseline incurs only the 13 false negatives and costs $C = 13$, cheaper than either system at its F1-selected point and cheaper than either unconstrained minimum; S2's minimum is in effect that baseline plus one residual wrong link. Asymmetric cost therefore needs a practice-grounded recall or false-link constraint rather than an arbitrary coverage cap, since an unbounded cost criterion on a NIL-dominated set is optimised by silence.

Every false positive received a warning, but entirely because R2 fired whenever the temporal check was unavailable. Only 54/450 records had a usable year literal and none of the 192 located works did. R1 never fired; for S1 this is partly structural, since conflicting candidates are vetoed before ranking. R3, the only warning that actually varies across decisions, fired for 19/34 S1 and 8/24 S2 false positives. The result is therefore an audit of evidence availability, not explanation quality. One error linked an anonymous copyist, *"Parmigianino (copia Da)"*, to the master.

6. Discussion

The staged results change how aggregate entity-linking outcomes should be read. A NIL prediction is often treated as one residual class, but N1, N2, and N3 have different remedies: expanding a knowledge base cannot fix a query that misses an existing item, and lowering a threshold cannot recover an item absent from the candidate set. Reporting only end-to-end precision and recall therefore conceals where intervention is needed and may wrongly attribute a retrieval failure to cautious model behaviour.

The profile rates are diagnostics, not estimates of social or corpus-level disparity, given the confounded partition. The zero rates for located works are nevertheless operationally important, identifying a

region where both target establishment and candidate availability are weak. For cultural properties that region also exposes a representational question, since a record may describe an object or attribution state for which Wikidata holds only a creator, a canonical work, or a holding institution. Failure to link is then not reducible to model accuracy but reflects a mismatch of descriptive granularities across infrastructures, which the staged protocol makes visible without determining its cause.

The trace audit yields a related lesson: warning coverage is meaningful only when the underlying evidence can vary across decisions. Here temporal insufficiency fires for every false positive because the field is almost always unavailable, while temporal counter-evidence is structurally suppressed by S1's veto, so a headline coverage rate of 100% overstates the practical content of the traces. A cataloguer-facing trace should distinguish conflicting, corroborating, and unavailable evidence, disclose when a check could not be performed, and cover relations central to cultural cataloguing: entity type, creator versus attributed master, copy or workshop relations, place, holding institution, and external identifiers. The copyist-to-master error shows how lexical similarity can preserve a name while reversing the relation the record asserts. The claim here is deliberately limited: not that traces improve professional decisions, but that a reproducible evaluation should preserve the provenance of each outcome and report evidence availability separately from link correctness.

7. Ethics, provenance, licensing

ArCo is CC-BY and Wikidata CC0; VIAF identifiers are stored only as identifiers. Fifteen records machine-flagged as possibly living persons are excluded from the released reference data. Because undated people cannot be screened reliably, a human release check must widen this exclusion where doubt remains. Reported figures use all 450 records.

Wrong links can reassign a record to the wrong person or object, and persistently unestablished targets may contribute to under-representation. These are risks motivating the evaluation, not harms measured by this study.

8. Limitations and next steps

The central limitation is the provisional reference from one non-version-pinned machine annotator: the precommitted codebook prevents retrospective adaptation but does not validate its judgements.

The immediate next step is a targeted human audit. Two annotators with cataloguing experience will independently review all 24 positive links, the deduplicated records implicated in either system's false positives, and stratified samples of 50 reference-NIL and 50 undecidable records, recording for each whether the proposed target is correct, whether a different target can be established, and whether the protocol should have abstained. Agreement will be reported per stratum with Krippendorff's α, disagreements adjudicated by a third annotator, and a corrected reference released where licensing and living-person screening allow. Profile rates, candidate recall, and linker estimates will then be recomputed, and N1 records will be separated into cases where expert review establishes a target missed by the original protocol and cases where no target is established after expert review.

The small positive set (13 on test) leaves wide intervals, the five profiles are confounded, the trace checklist was not evaluated with cataloguers, and the study covers one seeded sample from two ArCo classes whose targets differ in identity and granularity. Results are descriptive of this protocol and cannot yet support population, fairness, or harm claims.

9. Conclusions

The staged evaluation separates target establishment, candidate retrieval, link selection, and trace evidence through the N1, N2, and N3 conditions. Neither linker is adequate at the tested operating points: both assert wrong identifications on records for which the reference protocol established no target, both are dominated by always abstaining under an asymmetric cost, and their apparently

complete warning coverage rests on missing metadata rather than discriminating counter-evidence. The framework nevertheless identifies where an entity-linking pipeline loses epistemic support and why those losses matter for cultural-memory infrastructures. Human auditing is now the immediate requirement before the profile-conditioned rates are interpreted beyond this preliminary study.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude.ai, Grammarly for grammar and spelling checks, paraphrasing, and rewording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. Separately, as a research instrument rather than a writing aid, the reference annotation was produced by an LLM agent, as described in Section 3.

References

[1] C. Bianchini, S. Bargioni, C. C. Pellizzari di San Girolamo, Beyond VIAF, Information Technology and Libraries 40 (2021). doi:10.6017/ital.v40i2.12959.

[2] M. Conroy, Quantifying the gap: The gender gap in french writers' wikidata, Journal of Cultural Analytics 8 (2023). doi:10.22148/001c.74068.

[3] B. Patin, T. Youngman, E. Hauser, Toward an etiology of harm for knowledge organization: Onto-epistemic injustice in classificatory systems of record, Knowledge Organization 51 (2024) 495–513. doi:10.5771/0943-7444-2024-7-495.

[4] J. Baker, A. K. Mahal, "i have always found the whole area a minefield": Wikidata, historical lives, and knowledge infrastructure, International Journal of Digital Humanities 6 (2024) 217–236. doi:10.1007/s42803-024-00090-5.

[5] J. Lam, M. Cortina-Borja, R. Aldridge, R. Blackburn, K. Harron, Cluster-based name embeddings reduce ethnic disparities in record linkage quality under realistic name corruption: evidence from the north carolina voter registry, arXiv preprint arXiv:2601.07693 (2026).

[6] C. Barboi, F. Ouyang, L. Lembcke, A. Martin, A. Griffith, K. S. Allen, X. Li, H. Xu, S. J. Grannis, Evaluation of the accuracy of probabilistic record linkage across sociodemographic categories in 4 databases: Exploratory study, JMIR Formative Research 10 (2026) e78622. doi:10.2196/78622.

[7] T. Brasileiro Araújo, V. Efthymiou, K. Stefanidis, Fairness and explanations in entity resolution: An overview, IEEE Access 13 (2025) 145127–145143. doi:10.1109/ACCESS.2025.3599990.

[8] S. Vadrevu, R. Nagi, J. Xiong, W.-m. Hwu, xER: An explainable model for entity resolution using an efficient solution for the clique partitioning problem, in: Proceedings of the First Workshop on Trustworthy Natural Language Processing (TrustNLP), 2021, pp. 34–44. doi:10.18653/V1/2021.TRUSTNLP-1.5.

[9] H. Dong, J. Chen, Y. He, Y. Liu, I. Horrocks, Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), 2023, pp. 452–462. doi:10.1145/3583780.3615036.

[10] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents, in: Working Notes of CLEF 2022, 2022. doi:10.5281/zenodo.6979577.

[11] K. L. Polley, V. T. Tompkins, B. J. Honick, J. Qin, Named entity disambiguation for archival collections: Metadata, wikidata, and linked data, in: Proceedings of the Association for Information Science and Technology, volume 58, 2021, pp. 520–524. doi:10.1002/pra2.490.

[12] A. Fagerving, Wikidata for authority control: sharing museum knowledge with the world, in: Digital Humanities in the Nordic and Baltic Countries Publications, volume 5, 2023, pp. 222–239. doi:10.5617/dhnbpub.10665.