Evidence Before Generation: A Provenance-Aware Audit of Graph-Assisted Retrieval for Taiwanese Ethnohistory

Chun-Shih Cheng[1], **Yu-Jie Tsai**[1], and **Hsiao-Yu Wang**[2]
[1]Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung, Taiwan
t2025017@cyut.edu.tw, emcaut@gmail.com
[2]Department of Electrical Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan
shon0808@gmail.com

Abstract

Historical contracts, images, maps, and policy records document interactions among Indigenous peoples, migrants, tenants, officials, and institutions. Ordinary graph edges, however, obscure whether a relation is source-stated, model-inferred, or assigned during cataloguing. We present a provenance-aware resource audit and task formulation for Taiwanese ethnohistory organized around three research questions. For representation (RQ1), we separate event claims from time-indexed group-membership assertions and implement a deterministic evidence gate that either cites or abstains. On 68 public catalogue titles and seven public records of contracts from 1804–1818, the gate answers 44 of 68 catalogue event queries and abstains on the other 24, 18 of them because the counterparty is absent from the title. On 27 membership questions from the contract records, co-location and surname propagation, two inference patterns the schema prohibits, would attach 12 and 5 group labels that the records do not state. The gate attaches none, but it recovers only 9 of the 15 stated memberships because its attachment rule misses labels that govern a coordinated list of names. For pre-generation defects (RQ2), the catalogue audit finds 11 uncertain titles and five duplicate-title pairs, and the audit of a 46-node GraphRAG-oriented pilot graph finds 19 weakly connected components, 12 isolated nodes, three dangling source references, and two missing files. For retrieval (RQ3), a case-level diagnostic on five pre-authored questions shows that graph expansion helps one cross-label alias query and hurts one map query. The contribution is a resource audit with an executable gate, not a benchmark for extraction or generation.

Keywords: historical information extraction; graph-assisted retrieval; provenance; Taiwanese ethnohistory; Indigenous data governance

1 Introduction

Historical ethnic relations are rarely recorded in a single homogeneous corpus. A land contract may name a seller and a buyer; a catalogue may normalize its date; a map may locate a place name that is no longer in use; and a policy page may apply a contemporary group name. Integrating these sources is attractive for information extraction (IE), knowledge graphs, and retrieval-augmented generation (RAG), but it is also risky. A system may treat a catalogue description as contract text, equate a historical administrative label with contemporary identity, or infer group membership from a surname or residence. The resulting network can appear coherent while remaining unsupported by the documentary record.

GraphRAG enables structured retrieval over entities and relations (Edge et al., 2024). For historical inquiry, however, the appropriate unit is not an unqualified triple. A researcher needs to know which source states a relation, which span supports it, whether a place name is valid for the relevant period, and whether an ethnonym reflects source terminology, scholarly interpretation, or community self-identification. Merely producing a citation is insufficient: retrieval-augmented output may contain unsupported claims (Niu et al., 2024), while citation-generating models may cite evidence that does not fully support a claim (Gao et al., 2023).

This study addresses three research questions. **RQ1 (representation):** How should documented event relations, ethnonym mentions, group-membership assertions, evidence spans, and sources be represented without conflation, and can an evidence gate over this representation be executed to block unsupported group attribution? **RQ2 (auditing):** Which data and graph defects can be detected automatically before any generative model is evaluated? **RQ3 (retrieval):** How does simple graph-assisted expansion change the sources retrieved for the existing pilot questions, and why? Section 5.5 and Table 5 map each question to the evidence that answers it.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 393–401
November 6-7, 2026 ©2026 Association for Computational Linguistics

We analyze two complementary but separate artifacts. The first is the public catalogue of 68 Hsinchu Guanxi Regional Documents, including land and private contracts associated with early interactions between Plains Indigenous communities and Han settlers (Institute of Taiwan History, Academia Sinica, n.d.-a). The second is a manually curated GraphRAG-oriented pilot artifact assembled from public archival, image, map, policy, and bibliographic metadata. This artifact is a static demonstration that loads CSV records into a browser interface; it contains no live large language model (LLM), embedding retriever, or graph database. We neither merge the two artifacts nor claim that the pilot graph was sampled from the 68-document catalogue. A third, small set of public contract records is used only to exercise the evidence gate.

The paper is therefore a resource audit and task formulation with an executable gate. It does not report a deployed GraphRAG system, an extraction model, or answer generation. It makes three contributions. (1) A claim-centric representation that separates documented events, time-indexed membership, and mere ethnonym mention, together with a deterministic evidence gate that is implemented and executed on public records (Sections 4.1–4.2, 4.5, and 5.4). (2) Scripted audits of a public catalogue and a pilot graph that quantify provenance, connectivity, and leakage defects before modeling (Sections 4.3 and 5.1). (3) A case-level retrieval diagnostic and a mapping from the observed failures to annotation, evaluation, and governance requirements for a future benchmark of Chinese-language historical sources (Sections 4.4, 5.2, and 6).

2 Related Work

2.1 Historical Information Extraction

Historical IE must contend with optical character recognition (OCR) noise, orthographic variation, unstable names, and evidence that extends beyond a sentence. HistRED provides document-level relation annotations over Korean–Hanja historical records and multiple context lengths (Yang et al., 2023). CHisIEC spans 13 dynasties and supports named-entity recognition and relation extraction for ancient Chinese history (Tang et al., 2024). For more recent Chinese, Blouin et al. (2024) provide named-entity recognition and entity-linking data from historical newspapers dated 1872–1949. Fang et al. (2025) further show that language change and genre variation in sources from 1900–1950 materially affect Chinese word segmentation, part-of-speech tagging, and named-entity recognition.

HIPE-2026 extends historical IE to temporally grounded person–place relations and evaluates predictive accuracy, efficiency, and cross-domain generalization (Opitz et al., 2026). Its distinction between whether a person was ever at a place and whether the person was there at the document time illustrates why an atemporal person–place edge is insufficient. Our setting adds contractual roles, source-level provenance, spatial uncertainty, and the more sensitive question of group membership.

2.2 Graph-Assisted Retrieval, Attribution, and Cultural Heritage

RAG constrains generation with retrieved evidence (Lewis et al., 2020), while GraphRAG adds graph indexing and community-level summaries for corpus-level reasoning (Edge et al., 2024). Cultural-heritage research similarly combines ontological engineering, LLM-based extraction, and expert supervision. Schimmenti et al. (2026), for example, represent evidence and competing scholarly hypotheses in a knowledge graph. General provenance models distinguish entities, activities, and agents involved in producing information (Missier et al., 2013). These approaches support traceable access, but none implies that every extracted edge is historically warranted. Citation correctness and claim-level support remain distinct evaluation targets (Gao et al., 2023; Niu et al., 2024).

Accordingly, our pilot uses a graph only to retrieve candidate evidence and treats every relation as a reviewable claim. Source records and evidence spans are admission conditions for a claim, not decorative metadata. A model output may propose a candidate; it cannot serve as its own evidence.

2.3 Governance of Identity-Sensitive Data

Public availability does not imply unrestricted reuse or identity inference. The Collective Benefit, Authority to Control, Responsibility, and Ethics (CARE) Principles frame Indigenous data governance around those four dimensions (Carroll et al., 2020). We represent a historical ethnonym as terminology used by a source, not as an automatic determination of a contemporary person, family, or community. A surname, place of residence, transaction partner, or map co-location is insufficient to assert group membership. Even complete technical provenance does not replace community review or archival policy.

3 Data and Evidence Boundaries

3.1 The Guanxi Catalogue

The Institute of Taiwan History identifies series T0448D0369 as 68 Hsinchu Guanxi documents dated 1825–1923. The series includes land records, private contracts, Zhang-family partition records, and materials associated with descendants of Wei Agui; the collection overview explicitly mentions Plains Indigenous–Han interactions in land development (Institute of Taiwan History, Academia Sinica, n.d.-a). Item pages provide persistent identifiers, titles, dates, and access links, although many originals require an application or on-site consultation. We use all 68 public metadata records but do not treat catalogue titles as contract text.

A deterministic title lexicon assigns document-form categories; it does not extract ethnic relations. A year is normalized only when a title contains a reign-year expression or Gregorian year. Box placeholders and ellipses count as title uncertainty. Duplicate diagnosis removes the persistent item number before exact title-string comparison.

Metadata document form	n	%
License or permit	35	51.5
Inheritance partition or title	9	13.2
Private or marriage record	7	10.3
Payment or receipt	6	8.8
Sale contract	4	5.9
Tenancy or cultivation	3	4.4
Mortgage or other	4	5.9
Total	68	100.0

Table 1: Catalogue audit by title-derived document form. The “other” portion of the final category includes one contract certificate and one supplementary-payment record.

3.2 The Cross-Source Pilot Artifact

The second artifact defines seven annotation labels—`PLACENAME`, `ETHNICNAME`, `PERSON`, `EVENT`, `TEMPORAL`, `SOURCE`, and `EVIDENCESENTENCE`—whereas the audited node table instantiates four node categories: seven ethnonym terms, 13 places, 11 events, and 15 sources, for 46 nodes in total. It also contains 33 directed edges, 10 historical-to-modern toponym candidates, and five manually authored question cards. The source nodes comprise three Taiwan Memory image records, three Institute of Taiwan History archival records, three historical-map records, one additional artifact-specific historical record, three policy or local-history records, and two bibliographic records. Taiwan Memory and the Institute of Taiwan History provide heterogeneous image, contract, and catalogue metadata, while the historical-map service supports layer comparison (National Central Library, n.d.-a; Institute of Taiwan History, Academia Sinica, n.d.-b; Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, n.d.).

The graph’s own source-node table defines the scope of this audit; databases named only in planning documents are not counted. The five question cards are demonstration cases with author-specified expected source identifiers, not an independently sampled question-answering test set. Their single confidence field is an uncalibrated design score and is never interpreted as probability or answer accuracy.

3.3 Public Contract Records for the Gate

To exercise the gate on contract wording rather than titles, we use seven public records dated 1804–1818. Six are contract passages reproduced in two entries of the Encyclopedia of Taiwan Indigenous Peoples (Liao, n.d.; Encyclopedia of Taiwan Indigenous Peoples, n.d.): a petition (1804), a land sale (1805), two cultivation grants (1806, 1807), an irrigation agreement (1808), and a further grant (1818). The seventh is the Taiwan Memory catalogue record of the same 1818 grant (National Central Library, n.d.-b). We use the excerpts displayed on the public pages, not full transcriptions. These records lie outside series T0448D0369 and are not used in the retrieval diagnostic.

3.4 Source Tiers

Before extraction, each input is assigned a provenance tier: L1, a primary image or diplomatic transcription with a stable location; L2, institutional catalogue metadata; L3, a curatorial or scholarly page; and L4, a computational output. Within L3 we record whether a span reproduces an earlier contract passage (L3R) or is the page author’s own narrative (L3N); an entry-level ethnic classification (L3C) is stored as a scholarly classification, never as record-level membership evidence. These tiers describe derivation and access, not a universal truth ranking. The 68-record audit operates at L2. Pilot-graph edges are candidates derived from L2 or L3 material. L4 output may rank or propose a claim but cannot validate one.

4 Provenance-Aware Claim Graph

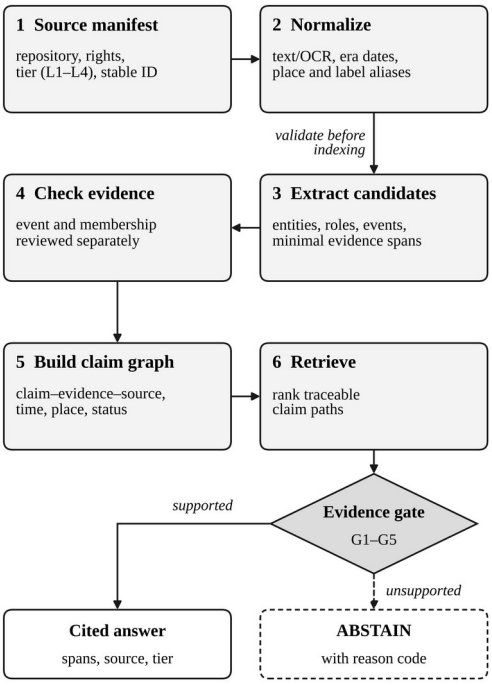

Figure 1: Provenance-aware workflow. Validation precedes graph indexing; the evidence gate (conditions G1–G5, Section 4.2) returns either a cited answer with its source tier or an abstention with a reason code.

4.1 Events and Membership Are Distinct Assertions

We propose the following target schema in response to the audit; its time-indexed membership field and separated status vocabularies are design revisions relative to the pilot artifact. For actors *u* and *v*, an event claim is:

$$c = \langle u, r, v, \tau, \ell, e, s, \pi, \rho \rangle \quad (1)$$

Here, *r* is a directed relation; τ and ℓ are normalized time and place; *e* is a minimal supporting evidence span; *s* is a source pointer with its tier; π is claim status (`CANDIDATE`, `VERIFIED`, `CONFLICTING`, or `UNSUPPORTED`); and ρ is review status (`UNREVIEWED`, `NEEDS_REVIEW`, `APPROVED`, or `REJECTED`). Seller–buyer, landlord–tenant, petitioner–authority, and witness–transaction roles must be preserved because reversing an edge changes the event.

Group membership is stored separately and is explicitly time-indexed:

$$m = \langle u, G, \tau_m, e_m, s_m, \pi_m, \rho_m \rangle \quad (2)$$

Here, *G* may be a named social group, a period-specific source classification, or `UNKNOWN`, and τ_m records its temporal scope. If a source merely mentions an ethnonym, the system may create a lexical node but may not assign that label to nearby persons. Aggregation from actors to groups is permitted only when membership at both endpoints has cited evidence or expert validation. The answer-level action is separate from claim status: the system either returns a cited answer or `ABSTAIN`. Thus, an actor-level event may be retained even when group membership cannot be supported.

4.2 Reified Provenance and the Evidence Gate

Figure 1 specifies a hard evidence gate. Rather than storing only $u \xrightarrow{r} v$, a claim node links the subject, object, relation type, time, place, evidence span, source record, claim status, and review status. An evidence span points to a source record, which retains its repository, persistent identifier, access tier, and rights status. Historical-label normalization preserves both the surface form and one or more reviewed candidates. Historical places map to candidate modern locations with temporal intervals and spatial uncertainty instead of being silently overwritten by present-day administrative units.

The gate returns `ABSTAIN` under five conditions: (G1) the evidence is empty or contains no label attached to the queried actor; (G2) a required actor is unresolved or unreadable; (G3) a membership assertion relies only on surname or co-location; (G4) the retrieval path contains only L4 outputs; and (G5) the only supporting relation is generic, such as `RELATED_TO`. Generic relations may support exploration but cannot alone support a historical answer. The gate is implemented as a deterministic module (Section 4.5). Because no generator is attached, it is evaluated as an answer-admission filter rather than as an intervention on generated text.

4.3 Artifact Audit

The audit tests: (i) required fields and non-empty values; (ii) whether both endpoints of each edge resolve to nodes; (iii) whether each edge's source identifier resolves to a source node; (iv) evidence and review-state completeness; (v) weakly connected components and isolated nodes; (vi) uncertainty and review status for toponym candidates; and (vii) the existence of files referenced by question cards. All tests are scripted and run without manual inspection. They do not establish historical truth; they identify conditions under which a provenance path cannot be reproduced.

4.4 Retrieval Diagnostic

We compare two deterministic source rankers over the 15 source nodes. The lexical ranker computes Jaccard overlap between the character-bigram set of a question and that of each source label plus note. The graph-assisted ranker adds the

artifact-provided heuristic weight of every edge attributed to source *s* whose subject or object label occurs verbatim in the question:

$$S(s \mid q) = J(B(q), B(t_s)) + \Sigma_{e \in E(q,s)} w_e \qquad (3)$$

Here, t_s is source text, $B(\cdot)$ is the character-bigram set, and w_e is the provided, uncalibrated edge weight. $E(q,s)$ is the set of edges whose source attribute equals *s* and for which at least one non-source endpoint label occurs verbatim in *q*. Edge direction is ignored only for label matching, and each qualifying edge is counted once even if both labels match or a label occurs repeatedly. Ties are broken by source identifier. For evaluation, the semicolon-delimited source field of each question card is parsed; only identifiers present in the source-node table are retained, while filenames such as `kg_edges.csv` are excluded. The diagnostic therefore concerns ranking within the closed 15-source inventory, not recovery of missing resources. For each question we record which expected sources enter the top three under each ranker and trace any change to the edges responsible. Because unnormalized edge-weight sums may favor higher-degree nodes and are not commensurate with a 0–1 Jaccard score, the two rankers are compared case by case rather than as competing systems.

4.5 Executable Gate Experiment

Event queries. A deterministic title grammar turns each catalogue title, as listed on the Institute's public exhibit pages, into a typed event claim: a title of the form "X gives Y a permit" yields issuer X and recipient Y, whereas "X issues a deed" yields only the issuing party. Sales, tenancies, mortgages, supplementary payments, and personal contracts require two parties; partitions require only the issuing branch. The seven contract records are encoded with the same schema, and each query asks which parties acted in which roles.

Membership queries. For every readable actor in every record—111 record–actor pairs, 84 from titles and 27 from contract records—a query asks which group the actor belonged to in that record; offices and institutions are excluded. We compare four configurations. S1 propagates a label to an unlabeled actor who shares a surname with a labeled actor elsewhere (leave-one-out). S2 assigns any label occurring in the same record (co-location). S3 combines an attached-label extractor with S2 and S1 without a gate. S4 passes the S3 candidates through the gate, which admits only a label syntactically attached to the actor—a community name or ethnonym, optionally followed by an office title, immediately before the name—in non-L4 evidence. S1 and S2 are not proposed as competitive baselines. They implement the two inference patterns that Section 2.3 rules out—group assignment from a shared surname and from co-occurrence in one document—so that the experiment can count how many labels each would attach that the record does not state.

Gold and metrics. The authors annotated the 27 contract-record pairs for whether the record explicitly attaches a group label to the actor, counting a label that governs a coordinated list of names as attached to every name in it; the 84 title pairs contain no group label, so any answer for them is unsupported. The attachment rule was written while inspecting 16 of the 27 pairs (development). The other 11 pairs come from contract text retrieved after the rule was fixed and serve as a held-out check. We report both the unsupported answers each configuration emits and the share of stated memberships it recovers. We define the Unsupported Ethnic Attribution Rate:

$$UEAR = N(U \cap M_{ans}) / N(M_{ans}) \qquad (4)$$

Here, M_{ans} is the set of non-abstained group-membership assertions and *U* is its unsupported subset. UEAR is reported with membership-answer coverage, the number of non-abstained membership answers divided by the number of membership questions; universal abstention would otherwise appear deceptively safe, and a zero denominator is reported as not applicable. Inter-group queries ask whether a member of one group acted toward a member of another and require the typed event and both endpoint memberships to pass.

5 Results

5.1 Catalogue and Graph Audits

Of the 68 catalogue records, 67 have a normalizable year (98.5%), and 11 titles contain uncertainty markers (16.2%). Five distinct title strings each form an exact duplicate pair, covering 10 records. Because these records may be copies, different originals, or catalogue variants, future data splits must group both exact and fuzzy title clusters. Licenses or permits account for 51.5% of the series (Table 1), creating a document-form coverage risk that future splits should measure and control.

Audit item	Observation
Nodes / directed edges	46 / 33
Edges with resolvable non-source endpoints	33 / 33
Edges with non-empty evidence and status fields	33 / 33
`CANDIDATE` / `NEEDS_REVIEW` status values	21 / 12
Weakly connected components / isolated nodes	19 / 12
Largest weakly connected component	16 nodes
Edges with a dangling source reference	3
Missing referenced intermediate files	2
Toponym uncertainty (low / medium / high)	2 / 7 / 1
Toponym candidates requiring further review	9 / 10

Table 2: Structural audit of the GraphRAG-oriented pilot artifact. Non-empty fields do not establish historical correctness.

All 33 directed edges have resolvable non-source endpoints, and their source, evidence, confidence, and review fields are non-empty. Field presence does not imply referential integrity or historical correctness. Across the recorded status values, 21 edges are marked `CANDIDATE` and 12 `NEEDS_REVIEW`; this conflation of claim and review state motivates the separated status spaces in Section 4.1. Three policy edges reference `POL_003`, which is absent from the 15 source nodes. Source identifiers are usually stored as edge attributes rather than connected source nodes, yielding 19 weakly connected components, 12 isolated nodes, and a largest component of only 16 nodes. Two intermediate files referenced by the question cards, `alignment.csv` and `retrieval.csv`, are missing; only `kg_edges.csv` is present.

The 10 toponym candidates are likewise provisional: two have low, seven medium, and one high spatial uncertainty; nine explicitly require geographic information system (GIS) or expert review. Each of the five question cards contains an answer, source list, evidence text, reasoning path, confidence, and limitation field. Schema completeness, however, does not demonstrate that an answer is entailed by a primary record.

5.2 Source Retrieval

Graph expansion leaves the top-ranked and top-five sources unchanged for all five question cards and changes the top three for two of them; on aggregate, macro Recall@3 moves from 0.750 to 0.767 while Recall@1 (0.483) and Recall@5 (1.000) are identical, which five questions cannot generalize. The two changed cases are informative. For the alias question connecting the historical term *Pepohoan* with a Chinese label, expansion through a candidate `SAMEAS` edge brings the third expected source into the top three. This edge should be understood as an unverified historical-label mapping, not identity equivalence. For the historical-map question, a bibliographic edge outranks two expected map sources and pushes one of them out of the top three. The graph therefore helps a cross-label query when a mapping edge exists and propagates source noise when an untyped bibliographic edge carries weight, which argues for typed, reviewable mapping claims rather than unrestricted expansion.

5.3 Source-Grounded Relation Cases

Table 3 stress-tests the representation using four Guanxi–Zhuqian events reproduced in a curated Indigenous encyclopedia entry (Liao, n.d.). These cases are not part of the 1825–1923 catalogue and were not used in retrieval evaluation. Their purpose is to show why event support and membership support require separate annotation.

Year	Documented event	Claim-level output
1804	Petition alleging intrusion, usury, coerced contracting, and property occupation.	`PETITION`; `CONFLICT`. Acts and collective labels remain source-attested.
1805	Seller bore a historical source label; buyer was explicitly labeled Han; Wei Fuxing witnessed.	`LAND_TRANSFER`; `WITNESS`. Preserve labels as source terms.
1806	Representatives granted land to cultivators explicitly labeled Han under rent conditions.	`CULTIVATION`. Preserve roles, rent, and collective labels.
1808	Landlord and named tenants set an irrigation route and annual water rent.	`WATER_RENT`. Tenant membership is unsupported; omit that assertion and abstain if asked.

Table 3: Hand-built claims from an L3 curatorial page reproducing contract passages. Event descriptions are paraphrases; historical labels are source terms, not contemporary identity determinations.

A generic `INTERACTS_WITH` edge would collapse petitioning, coercion, land transfer, cultivation, witnessing, and water rent. Conversely, assigning a group label to the 1808 tenant would create an intergroup edge not licensed by the reproduced passage. Equations 1 and 2 retain the event while allowing the answer policy to abstain from unsupported group aggregation.

5.4 Executable Evidence Gate

Event level. The gate answers 44 of the 68 catalogue event queries: 32 permits or rent receipts, nine partitions, two payment receipts, and one deed-tax certificate. It abstains on 24: 18 records whose second party does not appear in the title, including all 17 sales, tenancies, mortgages, supplementary payments, and personal contracts; five whose required party contains a box placeholder or ellipsis; and one untyped note. Catalogue titles alone therefore cannot support a relational answer for any bilateral contract in the series. All seven contract records pass, because their reproduced text names both required parties.

Config.	Titles	Contr.	Unsup.	UEAR	Stated
S1 Surname	5	6	10	0.91	1
S2 Co-location	0	27	12	0.44	10
S3 Ungated union	5	27	17	0.53	13
S4 Gated	0	9	0	0.00	9

Table 4: Non-abstained membership answers on the 84 title pairs (Titles) and 27 contract-record pairs (Contr.), unsupported answers, UEAR over all answers, and stated memberships answered with the correct label (Stated, of 15). S3 is the ungated union of the attached-label extractor, S2, and S1; S4 passes the S3 candidates through the gate. On the title subset, S2 and S4 give no answer, so their UEAR there is not applicable; all nine S4 answers come from contract records.

Membership level. Co-location answers all 27 contract-record questions, but 12 answers are unsupported, and in five cases it gives an Indigenous party the Han label that the record attaches to the other party. Surname propagation emits 11 assertions, of which 10 are unsupported: five Han labels for catalogue actors surnamed Chen, Han labels for three people surnamed Peng, and the community label for two witnesses surnamed Wei; its one correct label is a coincidence of surname. The gated configuration emits no unsupported label and abstains on 102 queries, 79 with no attached label and 23 for which only surname or co-location evidence existed. It recovers 9 of the 15 stated memberships: all 7 on the development pairs but only 2 of 8 on the held-out pairs. All six misses come from coordinated lists in the 1804 and 1806 records. There a single label governs several names, as in the grantees “Han Jiang Xianwei, Wu Yuanshu, Lin Huguan”, and the rule attaches it only to the first name; when the first name is preceded by an office title (*tongshi*) that the rule does not cover, it misses that name as well. The gate’s errors are thus omissions rather than unsupported attributions, and the held-out result shows that the attachment rule must handle coordinated lists before its coverage can be trusted.

Inter-group level. Six records have two personal parties. The gate answers the inter-group question for three of them (1805, 1807, and the encyclopedia reproduction of 1818), each supported by reproduced contract text (L3R). It abstains on the Taiwan Memory record of the same 1818 grant, whose catalogue metadata names the grantee only by role and omits the Han label stated in the contract; on the 1808 agreement, whose tenants carry a role rather than a group label; and on the 1806 grant, where the rule misses the grantors’ community label. A naive co-occurrence builder would add 62 untyped `RELATED_TO` edges for the same records; the gate admits none as sole support.

5.5 Answers to the Research Questions

RQ	Evidence	Answer
RQ1	Eqs. 1–2; §4.5; Tables 3–4	Events and memberships must be separate assertions. All seven contract records support an event answer, but only three of six support an inter-group answer; ungated heuristics would attach 10–17 unsupported group labels, while the gate attaches none and recovers 9 of 15 stated memberships.
RQ2	§4.3; Tables 1–2; §5.4	Defects are detectable by script before modeling: 11 uncertain titles, five duplicate pairs, a 51.5% permit skew, 24 of 68 titles that cannot support an event answer, three dangling references, 19 components, and two missing files.
RQ3	§4.4; §5.2	Expansion changes the top three for two of five questions: a mapping edge helps an alias query, and an untyped bibliographic edge hurts a map query.

Table 5: Research questions, supporting evidence, and answers.

6 Implications for a Benchmark

The present resources support schema testing and failure detection, not estimates of full-text extraction or answer generation. Each finding above maps to a requirement for a future benchmark of Chinese-language historical sources. Because titles cannot support relational answers for bilateral contracts (§5.4), annotation must start from rights-compatible images or diplomatic transcriptions with span coordinates. Because event and membership support diverge (Table 3, §5.4), annotators should mark them as separate layers, report agreement per layer, and have identity-sensitive labels reviewed by people with relevant community perspectives. Because

exact duplicates and a 51.5% permit share exist (§5.1), data splits must group title clusters, families, and contract chains, and results should be reported by document form. Because graph expansion propagated source noise (§5.2), retrieval should be compared with and without typed mapping claims and the gate, reporting citation precision alongside recall. Finally, UEAR (Eq. 4) should accompany relation F1 together with membership-answer coverage, since a gain in F1 cannot offset unsupported attribution.

Automation and human review. The audits and the gate run without human input, and their abstention reasons determine where human effort is spent. Abstentions caused by a missing counterparty (G2) should be routed to automatic full-text extraction once transcriptions exist, while only surname, co-location, or conflicting-label cases (G3) need identity-sensitive review; in Section 5.4 this triage would send 23 of 111 membership queries, rather than all 111, to reviewers, while the held-out misses identify where the automatic extractor must improve. Fuzzy duplicate clustering and candidate repairs for dangling references can likewise be proposed automatically. Under CARE, automation narrows the scope of human adjudication but does not replace it for identity-sensitive labels.

7 Ethics and Limitations

Archives are not neutral samples. Contracts overrepresent property, taxation, administration, and institutionally legible disputes; people without recognized rights or writing access may appear only through others' claims. Extraction counts therefore describe surviving records, not the full structure of social relations. Historical labels may have been externally imposed and may now be offensive. They should be retained only when evidentially necessary, marked as source terminology, and separated from contemporary names and self-identification.

The catalogue audit is small, regional, and metadata-only. Title-derived document forms require full-text confirmation. The pilot artifact has only 15 source nodes and five pre-authored questions whose expected sources were not independently adjudicated, and its edge weights are uncalibrated. The gate experiment uses short public excerpts rather than full transcriptions, a deterministic title grammar, and an attachment rule whose held-out recall is low. We do not evaluate generation, faithfulness, named-entity recognition, relation extraction, or historical truth. The two artifacts remain unintegrated, and the hand-built cases and contract records come from pages outside series T0448D0369.

The gate code, parsed catalogue titles, excerpt manifest, annotations, and per-query decisions are provided as supplementary material; the pilot artifact's schema, title lexicon, and audit scripts will be released with versioned resource snapshots.

Old documents may still be subject to archival terms, personal-exposure concerns, and community authority. Following CARE, future publication should include community consultation, tiered access, and correction or removal mechanisms rather than assuming that technical anonymization resolves governance.

8 Conclusion

Historical ethnic-relation inquiry is not merely a matter of linking names and places. Auditing 68 catalogue records reveals document-form skew, uncertain titles, and duplicate leakage before modeling, and auditing the pilot graph shows that apparently complete edges can coexist with isolated sources, dangling references, missing files, and unverified spatial alignments. Executing the evidence gate on public records shows that catalogue titles cannot support relational answers for bilateral contracts, that surname and co-location inference would attach group labels the records do not state, and that separating event from membership assertions lets a system answer what a source states while abstaining from what it does not. In retrieval, graph expansion helped where a mapping edge existed and hurt where an untyped edge carried weight. Abstention as a valid answer action, together with tier-aware citation, converts these limitations into testable requirements for future Chinese-language historical resources and evidence-constrained graph-assisted retrieval.

References

Baptiste Blouin, Cécile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 385–394, Torino, Italia. ELRA and ICCL. https://doi.org/10.18653/v1/2024.lrec-main.35

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1):43. https://doi.org/10.5334/dsj-2020-043

Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica. n.d. Taiwan Historical Maps. Accessed 10 August 2026. https://gis.rchss.sinica.edu.tw/twhgis/

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A Graph RAG approach to query-focused summarization. arXiv:2404.16130. https://doi.org/10.48550/arXiv.2404.16130

Encyclopedia of Taiwan Indigenous Peoples. n.d. Wei Shouzong. Accessed 5 October 2026. https://aborgpedia.alcd.center/detail?cat=0&id=8239&race=0&search=&writer=

Zhao Fang, Liang-Chun Wu, Xuening Kong, and Spencer Dean Stewart. 2025. A comparative analysis of word segmentation, part-of-speech tagging, and named entity recognition for historical Chinese sources, 1900–1950. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 1–6, Albuquerque, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlp4dh-1.1

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

Institute of Taiwan History, Academia Sinica. n.d.-a. Hsinchu Guanxi Regional Documents, series T0448D0369. Accessed 10 August 2026. https://archives.ith.sinica.edu.tw/collections_list_02.php?no=12

Institute of Taiwan History, Academia Sinica. n.d.-b. Taiwan Archival Information System: Collection search and access levels. Accessed 10 August 2026. https://archives.ith.sinica.edu.tw/resources_list.php

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Chih-hsuan Liao. n.d. Wei Fuxing. Encyclopedia of Taiwan Indigenous Peoples. Accessed 10 August 2026. https://aborgpedia.alcd.center/detail?cat=0&id=6876&race=0&search=&writer=

Paolo Missier, Khalid Belhajjame, and James Cheney. 2013. The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776. https://doi.org/10.1145/2452376.2452478

National Central Library. n.d.-a. Taiwan Memory. Accessed 10 August 2026. https://tm.ncl.edu.tw/about?lang=chn

National Central Library. n.d.-b. Cultivation grant of green upland issued by Wei Fuxing, former *tumu* of Zhuqian She (1818), item 2591675. Taiwan Memory. Accessed 5 October 2026. https://tm.ncl.edu.tw/article?u=015_002_0000426303

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.585

Juri Opitz, Maud Ehrmann, Corina Raclé, Andrianos Michail, Matteo Romanello, and Simon Clematide. 2026. Overview of HIPE-2026: Person–place relation extraction from multilingual historical texts. arXiv:2606.25935. https://doi.org/10.48550/arXiv.2606.25935

Andrea Schimmenti, Valentina Pasqual, Fabio Vitali, and Marieke van Erp. 2026. Knowledge graphs generation from cultural heritage texts: Combining LLMs and ontological engineering for scholarly debates. *Journal of Documentation*, 82(7):206–250. https://doi.org/10.1108/JD-07-2025-0203

Xuemei Tang, Zekun Deng, Qi Su, Hao Yang, and Jun Wang. 2024. CHisIEC: An information extraction corpus for ancient Chinese history. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 3192–3202, Torino, Italia. ELRA and ICCL. https://doi.org/10.18653/v1/2024.lrec-main.283

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.180