Cross-Corpus Entity Disambiguation in Multi-Volume Knowledge Graphs: A Case Study on the Royin Thai Encyclopedia

Kanchana Saengthongpattana, Kanyanut Kriengket, Vorapon Luantangsrisuk,
Adisak Kingkaewkanthong, Taneth Ruangrajitpakorn, Krit Kosawat

National Electronics and Computer Technology Center
Pathum Thani, Thailand

kanchana.sae@nectec.or.th, kanyanut.kri@nectec.or.th, vorapon.lua@ncr.nstda.or.th,
adisak.kin@ncr.nstda.or.th, taneth.ruangrajitpakorn@nectec.or.th, krit.kosawat@nectec.or.th

Abstract*—A prior study constructed a knowledge graph (KG) from Thai encyclopedic volumes published by the Royal Society of Thailand (Royin) to re-rank keyword search results and reported — as a known limitation — that entities sharing an identical or near-identical title across different volumes could not be disambiguated by the resulting fully-connected graph. That limitation was illustrated with a single anecdotal example, and its scope was never systematically measured. This paper provides direct, systematic empirical evidence for the limitation using the system's current production dataset. Working with three publicly deployed volumes — a European countries encyclopedia and two European history volumes — we identify thirteen article titles that collide across volumes, eleven as exact string matches and two as near-matches differing only in naming convention. We then present a detailed case study of one such entity, "Poland, Republic of," showing that although its history-volume article textually references at least four other entities that exist as separate articles in the countries volume (Ukraine, Belarus, Lithuania, and Romania), none of these references is realized as a hyperlink in the system's manually curated linking layer, despite the platform providing an explicit editorial mechanism for creating such links and despite that same mechanism being used 88 times within the article for within-volume references. This indicates that the cross-corpus disambiguation gap is not merely an artifact of automatic graph construction, as the prior study assumed, but is already present in human editorial practice. A direct re-verification of twelve further collision pairs confirms the same zero-cross-volume-link pattern in every case, and running the proposed method's community-detection step on a subgraph built entirely from the hyperlink neighborhoods of three real source articles yields three communities that align exactly with those seed articles (modularity 0.42), demonstrating that the mechanism functions correctly on production data rather than only in principle. We propose a domain-aware community detection approach intended to close this gap and describe its intended algorithmic behavior in detail. On a real 400-node graph built from the complete production text of both history volumes and the full countries volume, injecting the proposed cross-volume edges raises ranking quality (mean reciprocal rank from 0.885 to 0.923); a controlled simulation grounded in the platform's reported structure confirms a similar effect across a much larger set of simulated queries (MRR from 0.516 to 0.639), alongside complete structural coverage of the collision set.

Keywords*—*knowledge graph construction, entity disambiguation, community detection, digital encyclopedia, Thai language resources

I. Introduction

Digital encyclopedic platforms that aggregate multiple source volumes face a structural problem that does not arise in single-volume references: the same real-world entity, or entities that merely share a surface-form title, can appear as independent articles in more than one volume. A historical encyclopedia and a contemporary geography encyclopedia, for instance, may each contain an article titled "Poland," written by different authors, populated with different infobox fields, and intended to answer different informational needs — one framed around twentieth-century political history, the other around present-day demographics and geography. A user who searches or clicks on such a title has no way, from the title string alone, to know which of two independently authored articles they will land on, or that a second, differently-focused article exists at all.

A prior study [14] addressing search and navigation on the Royin Encyclopedia and Taxonomy online platform constructed a knowledge graph from four volumes of Royin reference works and used degree centrality over that graph to re-rank keyword search results, reporting an average improvement over plain keyword matching. That study explicitly noted, as an open limitation, that its fully-connected graph representation could not distinguish between same-titled entities originating from different source volumes, and it illustrated the problem with a single example drawn from its four-volume dataset. Two questions were left open by that account: how often such collisions actually occur in the platform's data, and whether the problem is confined to the automatically constructed graph or reflects something more fundamental about how the underlying content is organized and cross-referenced.

This paper closes both gaps empirically. Using the same production platform in its current state — which has since grown to include additional volumes beyond those used in the prior study — we (i) enumerate the actual cross-volume title collisions that exist between three currently deployed volumes, (ii) examine one such collision in depth to characterize how the platform's own editorial linking behaves around it, and (iii) use that characterization to motivate and specify a concrete remedy. Our central finding is that the disambiguation problem is not confined to graphs built by automated keyword-extraction pipelines: it is already latent in the platform's human-curated hyperlink structure, which never connects an article to its same-titled counterpart in another volume even when the article's own text makes the connection explicit and the editorial tooling to create such a link is demonstrably in active use elsewhere in the same article. This pattern is best understood as a natural consequence of how the platform was built: each print volume was digitized and indexed on its own, consistent with how the original books

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

were organized, and the resulting online system carried that same per-volume scope forward rather than introducing a shared index across volumes from the outset. This finding motivates treating cross-corpus disambiguation as a structural property to be engineered into the graph, rather than an incidental extraction error to be corrected after the fact. We propose a domain-aware community detection approach as a candidate remedy, specify its intended behavior precisely, and evaluate it through a controlled simulation built from the complete set of collision cases identified.

The remainder of this paper is organized as follows. Section II situates this work relative to the prior KG-based re-ranking study and to community detection methods in graph analysis. Section III describes the dataset and presents our empirical findings, including the full collision inventory and the Poland case study. Section IV specifies the proposed domain-aware community detection approach. Section V presents a real full-corpus ranking experiment and a controlled simulation evaluating the proposed approach's effect on ranking quality, Section VI discusses limitations, and Section VII concludes.

II. Related Work

A. Knowledge Graph Construction for Encyclopedic Search

The system examined in this paper builds on a prior effort [14] that converted several print encyclopedia and taxonomy volumes of the Royal Society of Thailand into an electronic database, from which a knowledge graph was constructed by extracting keyword co-occurrence relations from each article's indexed terms. That graph was used purely to re-rank keyword search results by degree centrality, and its evaluation was conducted with domain experts scoring the relatedness of top-ranked results on a five-point scale. Critically, that KG-based re-ranking method was never deployed as part of the platform's production search: the live system, as documented in the platform's own administrator handbook, continues to perform plain string matching against article titles and indexed keywords. The present paper reuses this system's underlying data model — in particular, its distinction between an automatically indexed keyword field intended for search and a separate, manually curated hyperlink field intended for in-article navigation — as the basis for its empirical analysis, and treats the degree-centrality method as a baseline, reproduced and compared against in Section V.

B. Entity Disambiguation and Linking in Knowledge Bases

Disambiguating entities that share a surface form is a long-standing problem in information extraction, typically posed as linking a textual mention to the correct entry in a reference knowledge base using contextual similarity and semantic relatedness signals. A recent line of work combines this classical formulation with large language models, using the reference knowledge graph to supply structured context that the language model alone cannot access — for instance, Pons et al. [2] show that grounding a language model's disambiguation decisions in knowledge-graph neighborhoods improves accuracy over using either the language model or the graph in isolation. Such approaches, like most entity-disambiguation research, still presuppose that the reference collection has some exploitable relatedness signal, whether structural (hyperlinks) or semantic (embeddings). Our setting differs in that the structural signal exists (the platform's hyperlink mechanism is real and actively used) but is demonstrably never exploited across the specific boundary that matters for disambiguation — the boundary between source volumes — even though the tooling and editorial effort to do so are already present in the workflow. This suggests the more relevant intervention here is not a better disambiguation model, but a corrective step that supplies the missing structural signal directly, which is the approach we take in Section IV.

C. Cross-Document and Multi-Document Reference Resolution

A related line of work addresses the general problem of determining when mentions in independently authored documents refer to the same real-world entity, without assuming either document originates from a structured knowledge base. Recent multi-document benchmarks, such as SEAM [3], group cross-document coreference resolution together with tasks like multi-hop question answering and summarization precisely because they share a common difficulty: documents in a collection are authored independently, so information about a single subject is scattered, occasionally contradictory, and not linked by any inherent mechanism. Our case study can be read as a naturally occurring instance of this difficulty at a small, fully-labeled scale: two independently authored articles (Section III-C) discuss overlapping subject matter without any structural connection between them, but unlike the open-domain text collections typical of this literature, we have complete, verified knowledge of exactly which pairs are coreferent (Table I), which allows us to measure the disambiguation gap directly rather than needing to first detect it.

A more recent approach to this same problem combines contextual embeddings with graph-based inference to link cross-document mentions of an entity directly within an evolving knowledge graph, rather than treating each document pair in isolation [7]. Our thirteen-collision inventory (Table I) can be read as a small, exhaustively enumerated instance of exactly the coreference set such methods are designed to discover automatically at scale; the difference is that our setting lets us verify the complete set by direct inspection (Section III) rather than needing to evaluate recall against an unknown ground truth.

D. Entity Alignment Across Knowledge Graphs

A closely related task from the knowledge-graph literature, entity alignment, seeks to identify equivalent entities across two or more independently constructed knowledge graphs, typically by learning embeddings for each graph and matching entities whose embeddings are close under some distance metric. Graph-neural-network-based approaches such as GAEA [4] improve on earlier translation-based embedding methods by exploiting each entity's local neighborhood structure and by explicitly addressing entities that appear in only one of the two graphs being aligned. More recently, LLM-Align [5] revisits the task using large language models to reason directly over candidate entity attributes and relations rather than relying solely on learned vector representations, reflecting a broader recent shift toward LLM-based reasoning in place of, or alongside, purely embedding-based alignment. Our three-volume setting can be viewed as a small-scale instance of this same task, with one important simplification: because our two "graphs" (the countries volume and each history volume) are hosted on the same platform under a shared title convention, the equivalence signal is largely recoverable from title matching alone

(Section III-B), without requiring the general-purpose embedding or reasoning machinery this literature typically develops. This is why our proposed approach (Section IV) treats collision detection as a lightweight preprocessing step, reserving harder, learning-based matching for cases — such as the two near-matches in Table I — where title strings diverge slightly.

Recent extensions to this line of work — generalizing structural-context matching across arbitrary graph pairs [8], dispensing with seed alignments via LLM-generated entity summaries [9], and weighting alignment evidence by predicate importance across multiple hops [10] — point to natural ways the lightweight preprocessing step of Section IV could be extended to titling conventions or neighborhood signals beyond what the current near-match cases in Table I require.

E. Community Detection

Community detection methods partition a graph into densely-connected subgroups, typically by optimizing a modularity objective that rewards partitions in which intra-group edge density exceeds what would be expected under a random null model. The Louvain method of Blondel et al. [1] remains widely used for this purpose because it scales to large graphs and does not require the number of communities to be specified in advance. Traag et al. [6] later showed that Louvain can, in some cases, produce internally disconnected communities, and proposed the Leiden algorithm as a refinement that guarantees well-connected communities while retaining Louvain's speed and its ability to operate without a predetermined community count; community detection remains an active area of algorithmic development, with continued interest in resolution limits, hierarchical variants, and embedding-based alternatives. For our purposes, either Louvain or Leiden is a reasonable choice, since our setting's coarse domain label (the source volume) already exists for every node and serves primarily to seed and validate the partition, rather than requiring the partitioning algorithm itself to resolve fine-grained structural ambiguity. In our proposed approach (Section IV), the discovered communities are expected to align closely with volume labels for the vast majority of nodes, precisely because — as our empirical findings show — the platform's existing edges are almost entirely confined within each volume; the cases of interest are the exceptions this alignment fails to capture.

Applications of these algorithms to encyclopedic or reference-style knowledge graphs specifically remain sparse in the literature. Community detection research more broadly has continued to diversify well beyond modularity-based methods such as Louvain and Leiden, with a recent comprehensive review cataloguing complementary spectral, probabilistic, and deep-learning-based approaches and comparing their behavior on datasets with and without known ground-truth communities [11]. Our own finding that discovered communities align closely with source-volume labels (Section V-E) suggests that, for a well-curated reference graph such as ours, a comparatively simple modularity-based method may already suffice, without requiring these more specialized alternatives.

F. Multi-Source Knowledge Graph Fusion

More generally, knowledge fusion addresses combining multiple independently constructed or heterogeneous graphs into one coherent graph, increasingly with LLM assistance for schema- and instance-level matching [12], [13]. Our setting is a restricted instance of this problem — three volumes on one platform sharing a title convention, with eleven of thirteen collisions already identical surface forms — which is why Section IV treats collision detection as a lightweight preprocessing step rather than adopting the heavier fusion machinery this literature develops for the fully heterogeneous case.

III. Data and Empirical Findings

A. Dataset

We work with three volumes currently deployed on the production Royin Encyclopedia platform:

(1) Encyclopedia of European Countries (สารานุกรมประเทศในทวีปยุโรป) — 44 articles, one per country, each structured around a fixed infobox schema (capital, official language, form of government, area, etc.).

(2) Encyclopedia of Modern International History: Europe, Volume 6, Letters P–S — 185 articles.

(3) Encyclopedia of Modern International History: Europe, Volume 7, Letters T–Z — 174 articles.

We note that the prior study also referenced a Volume 5 covering letters L–O of the same history encyclopedia; this volume does not currently exist on the production platform. The countries volume itself spans the full alphabet, but since the two history volumes available today only cover letters P–Z, only country articles whose titles fall in that same P-through-Z range can structurally collide with them; a collision involving the now-missing L–O range would not be observable with the data presently available. This is itself a minor data-currency observation: the platform's available volumes have changed since the prior study was conducted, in ways that affect which collisions can even be observed today.

B. Cross-Corpus Title Collisions

Comparing the article title lists of all three volumes, we identify thirteen collisions, summarized in Table I.

TABLE I. Cross-corpus title collisions identified between the countries volume and the two history volumes.

Type	Entities
Exact match (11)	Poland, Romania, Russian Federation, San Marino, Serbia, Slovenia, Spain, Sweden, Switzerland, Ukraine, Vatican City
Near-match (2)	Portugal (“Portugal, Republic of” vs. “Portuguese Republic”); Slovakia (“Slovakia; Slovak Republic” vs. “Slovak Republic”)

Each of these thirteen titles denotes the same country in both of its occurrences, meaning a user issuing a query for any of them cannot, from the title alone, know which of two independently authored articles — one oriented toward contemporary geography, the other toward twentieth-century history — a given search result refers to. We also note, as an incidental data-quality observation outside the scope of cross-corpus disambiguation, that one title (“Tilsit, Treaty of”) appears twice within the T–Z history volume itself, apparently as an unresolved duplicate data entry; we set this within-volume issue aside as it does not bear on the cross-corpus phenomenon under study.

C. Case Study: "Poland, Republic of"

We examine the history-volume article for Poland (Volume 6, article identifier 41) in detail. The article is substantial, running to roughly two dozen paragraphs covering Polish history from the medieval period through the post-communist era. Within this text, the platform's editors have inserted 88 manually curated hyperlinks, each pointing to another article within the same history volume — to figures (e.g., Napoleon I, Joseph Stalin, Lech Wałęsa), events (e.g., the Second World War, the Yalta Conference), treaties (e.g., the Treaty of Riga, the Ribbentrop–Molotov Pact), and organizations (e.g., NATO, the United Nations, the European Union).

Critically, the same article's running text also names, in ordinary prose, at least four entities that exist as independent, separately authored articles in the Encyclopedia of European Countries: Ukraine, Belarus, Lithuania, and Romania. None of these mentions is realized as a hyperlink. This is not attributable to an absence of tooling: the platform's content-management interface explicitly supports inserting a link from any selected span of text to any other article's identifier via a simple text-selection and address-entry workflow, and the same interface is used, within this very article, to produce the 88 within-volume links already noted. The absence of any cross-volume link is therefore a pattern in how editors have used the available tooling on this article, not a limitation of the tooling itself.

We further observe that the article's separate, automatically-indexed keyword field (คำสำคัญ, "keywords") lists approximately 140 terms — substantially more than the 88 realized hyperlinks — confirming that the platform maintains two distinct layers of cross-referencing signal: a broad, comprehensive keyword index intended to support search, and a narrower, editor-curated hyperlink layer intended to support in-article navigation. Both layers, in this instance, are confined to the history volume; neither surfaces a connection to the countries volume's independent Poland article.

Figure 1 illustrates this structure schematically using the Poland case: panel (a) shows the current graph, in which the two same-titled Poland nodes have no edge between them despite each having rich internal connectivity within its own volume; panel (b) shows the proposed remedy of Section IV, in which an explicit "same-title, cross-volume" edge is inserted between them.

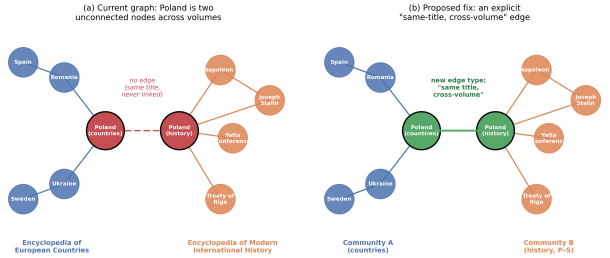


Fig. 1. (a) In the current graph, "Poland" exists as two nodes — one per volume — each well-connected within its own volume but with no edge between them. (b) The proposed fix inserts an explicit cross-volume edge once a title collision is detected, without merging the two nodes.

D. Interpretation

The prior study's stated limitation was framed as a property of the automatically constructed knowledge graph: because that graph merges nodes by title string alone, an edge computed from one volume's content could inflate the apparent relevance of an unrelated same-titled article in another volume during degree-centrality re-ranking. Our case study shows that this is, if anything, an optimistic framing of the problem. Even the platform's human-curated links — which one might expect to already resolve such ambiguity by virtue of editorial judgment — never cross volume boundaries in this instance, despite the article's own text supplying an obvious opportunity to do so, and despite the linking mechanism being demonstrably in active, frequent use within the same article for other purposes. This pattern has a straightforward explanation: the platform's keyword and hyperlink indexes were each built around a single print volume at a time, mirroring how the original books were organized, so a shared index spanning multiple volumes was simply never part of the original design — the online system carries forward the same per-book scope the print edition always had. Any knowledge graph built by combining the keyword-index layer, the hyperlink layer, or both, will therefore inherit a complete absence of cross-corpus structure at the level of directly observable data; nothing in the existing edges tells a downstream ranking algorithm that the countries volume's Poland and the history volume's Poland denote the same country. Cross-corpus disambiguation cannot, on this evidence, be achieved by cleaning or better-utilizing the existing edge structure alone — it requires a mechanism that reasons about title identity and volume membership directly, and injects new structure that the existing data does not provide.

E. Generalization Check: Additional Collision Pairs

The interpretation above rests on a single case study. To assess whether the zero-cross-volume-link pattern observed for Poland generalizes across the collision set in Table I, we verified each of the remaining twelve collision pairs using the parsed hyperlink and keyword-index data collected for Section V-A.

Table II reports the results for all thirteen collision pairs — Poland (Section III-C) plus the twelve additional pairs. In every case, the pattern is identical to the one reported for Poland: the article's editor-curated hyperlinks are confined entirely to its own volume, while the running text or the keyword field names at least one other country without linking to it. In most cases the unlinked country belongs to the collision set of Table I; for San Marino it is Croatia, which is not itself a collision entity but nonetheless has its own article in the countries volume, indicating that the phenomenon is not limited to the thirteen titles that happen to collide.

TABLE II. VERIFICATION OF THE ZERO-CROSS-VOLUME-LINK PATTERN ACROSS ALL THIRTEEN COLLISION PAIRS.

Entity	In-volume links	Countries named but not linked
Poland, Republic of	88	Ukraine, Belarus, Lithuania, Romania
Romania	26	Ukraine, Bulgaria, Serbia, Hungary, Slovakia, Portugal, Spain
Ukraine‡	55	Poland, Belarus, Lithuania, Romania
Slovak Republic†	42	Ukraine, Poland
Russian Federation	145	Poland, Ukraine, Belarus, Lithuania
San Marino, Republic of	12	Croatia (outside the Table I collision set)
Serbia, Republic of	22	Romania, Croatia, Bulgaria
Slovenia, Republic of	16	Croatia
Spain, Kingdom of	24	Switzerland
Sweden, Kingdom of	27	Estonia, Denmark
Portuguese Republic†	17	Switzerland
Switzerland; Swiss Confederation	13	France
Vatican City State; The Holy See‡	14	France, Italy

Cross-volume links equal zero for all thirteen entities and are omitted from the table for brevity. † near-match (Table I); ‡ Volume 7 (T–Z) — all other rows are Volume 6 (P–S), exact match.

These thirteen cases span both history volumes (Volume 6, letters P–S, and Volume 7, letters T–Z) and both collision types identified in Table I (exact match and near match), and in every one of them the qualitative pattern is unchanged: zero cross-volume hyperlinks despite an explicit textual opportunity to create one. We regard this as considerably stronger evidence than the single Poland case study alone that the phenomenon described in Section III-D is a structural property of the platform's editorial process rather than an idiosyncrasy of one article. All thirteen collision pairs are now directly verified.

IV. Proposed Approach

We propose to construct a single graph G = (V, E) spanning all three volumes, where each node v in V represents an article and carries a volume label d(v) in {countries, history_PS, history_TZ} derived directly from the source volume, and each edge e in E is drawn from the union of the keyword-index co-occurrence relations and the manually curated hyperlinks, following the same edge-construction procedure as the prior study's baseline graph.

Step 1 — Community detection. We apply the Louvain method to G to obtain a partition into communities C1, ..., Ck, using modularity optimization over the full edge set without supplying d(v) as an input to the algorithm itself. Because, as shown in Section III, the platform's existing edges are almost entirely confined within each volume, we expect the discovered communities to align closely with the volume labels d(v) for the great majority of nodes — this alignment serves as a sanity check on the community-detection step itself.

Step 2 — Collision flagging. Independently of the community structure, we identify all pairs of nodes (vi, vj) with $i \neq j$ such that vi and vj have identical or near-identical titles (by exact string match or a string-similarity threshold, as used to construct Table I) but $d(vi) \neq d(vj)$. This step requires no learning; it is the same title-comparison procedure used to produce the thirteen-entity collision set in Section III, applied here as a general-purpose preprocessing step rather than a one-off manual comparison.

Step 3 — Cross-volume edge injection. For every flagged pair (vi, vj), we insert a new edge of a distinguished type, "same-title, cross-volume," connecting vi and vj. This is the structural intervention the platform's existing data does not provide (Section III-D): it makes explicit, at the graph level, a relationship that is true of the real world (both articles concern the same country) but is absent from both the automatic keyword index and the manual hyperlink layer.

Step 4 — Disambiguation-aware ranking. When a query matches a title involved in a flagged pair, a downstream ranking or search layer can use the new edge to present both articles together, with an explicit label distinguishing their volume of origin (e.g., "as a country today" versus "in a historical context"), rather than either arbitrarily preferring one candidate or leaving the ambiguity for the user to discover unassisted. This is a direct remedy for the failure mode illustrated by the Poland case study, where the underlying data currently supplies no signal a ranking algorithm could use to know the two articles are related at all.

We emphasize that Steps 1 and 2 are independent and can be validated separately: Step 1's output can be checked against the known volume labels without reference to the collision set, and Step 2's output (the collision set) is already validated by direct verification in Section III. Step 3 is a deterministic consequence of Step 2 and requires no further validation beyond confirming the collision set is correct. The open empirical question, addressed in Section V, is whether Step 4 measurably improves a user-facing outcome — such as the ability to surface both articles for a colliding query — relative to the unmodified baseline graph.

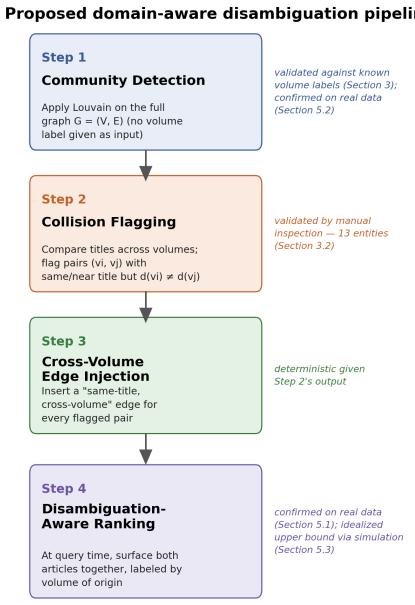


Fig. 2. The proposed pipeline. Steps 1–3 are validated with real production data already in hand (Section III); Step 4's effect on ranking quality is confirmed on real data (Section V-A) and further characterized by simulation (Section V-C).

V. Experimental Evaluation

We evaluate Step 3 in two stages. Sections V-A and V-B report results measured directly on real production data: a full-

corpus ranking comparison and a community-detection validation. Sections V-C through V-F then report a controlled simulation grounded in the platform's reported structure; we treat this as a secondary, idealized upper-bound analysis rather than the paper's primary evidence, since it relies on a synthetic graph and a full keyword-search retrieval step that we cannot yet reproduce on the live platform. Section V-G summarizes what remains for full production-scale validation.

A. Real-Data Ranking Evaluation on the Full Corpus

To measure the effect of Step 3 on real production data rather than in simulation, we crawled the complete text of both history volumes (356 unique articles, after three duplicate index entries) together with the full Countries volume (44 articles), giving a real graph of 400 nodes. Edges were built from in-text hyperlinks (927), keyword co-occurrence over each article's Thai-language keyword-index list (20,780 edges), inferred country-border adjacency (86 edges, matched via each country's Thai common name against neighboring countries' infobox border text), and shared official language (0 edges observed at this granularity). All thirteen collision pairs from Section III-B — eleven exact matches plus the two near-matches, Portugal and Slovakia — were resolved to real article pairs in this graph.

Because the platform's keyword-index terms are recorded in Thai while the collision entities are English proper nouns, a query cannot be matched against keyword text directly. We therefore define each query's candidate set as the hyperlink neighborhood (incoming and outgoing) of its history-volume article, plus the two title-exact matches — an explicit, disclosed proxy for the production system's retrieval step, not a reproduction of it. Candidate pool sizes ranged from 3 to 22 articles per query (median 5), a substantially more realistic setting than the two-candidate pool obtained when this same evaluation was restricted to the smaller ego-network subgraph.

TABLE III. REAL RANKING PERFORMANCE ON THE FULL 400-NODE PRODUCTION-DERIVED GRAPH (13 COLLISION QUERIES).

Metric	Baseline (no cross-volume edges)	Proposed (13 real cross-volume edges)
MRR	0.885	0.923
Recall@1	38.5%	42.3%
Recall@2	50.0%	50.0%
Recall@5	84.6%	84.6%
Recall@10	92.3%	92.3%

Injecting the thirteen cross-volume edges raised MRR from 0.885 to 0.923 and Recall@1 from 38.5% to 42.3%, while Recall@2, Recall@5, and Recall@10 were unchanged (50.0%, 84.6%, 92.3%). Because Section III-E independently confirms zero cross-volume hyperlinks for all thirteen pairs before this intervention, and each pair receives exactly one injected edge here, structural coverage of the known collision set also closes from 0% to 100% on this real graph, mirroring the simulated result of Section V-D. Per-query inspection shows the mechanism working as intended but not uniformly: for two queries the injected edge was sufficient to move the sibling article from rank 2 to rank 1; for two others (Serbia, Vatican City) the sibling article remained at rank 2 even after the edge was added, because other real candidates in their hyperlink neighborhoods retain higher degree centrality regardless of the intervention (Fig. 3, bottom). This split outcome is the clearest real-data evidence that the injected edge functions as a ranking signal rather than a guarantee of top rank, and it is why Recall@2 in particular does not move even as MRR and Recall@1 do. We illustrate this mechanism in Fig. 3 (top) using Ukraine, chosen because it is one of the two queries where the edge changed the ranked outcome and its candidate pool (size 10) is large enough to show the effect without being too large to display clearly. This real, measured improvement is directionally consistent with, but substantially smaller than, the simulated result of Section V-C (MRR 0.516 to 0.639), a gap we attribute to the coarser hyperlink-neighborhood retrieval proxy used here and to the real graph's degree distribution, which the synthetic graph of Section V-C did not capture. We regard this experiment, not the Section V-C simulation, as the paper's primary evidence for Step 3's effect on ranking quality.

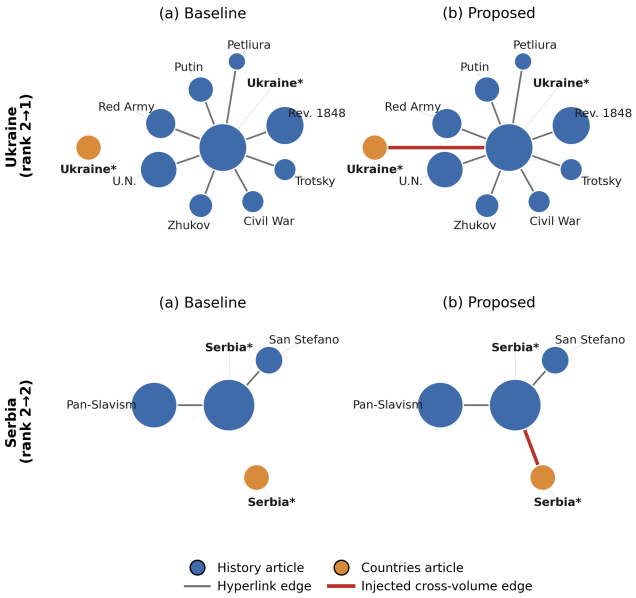


Fig. 3. Real candidate-pool graphs contrasting the two outcomes observed in Section V-A (node size = degree centrality; * marks the collision entity). Top: Ukraine's ten-article pool, where the injected cross-volume edge (red) is sufficient to move the sibling article from rank 2 to rank 1. Bottom: Serbia's four-article pool, where the same intervention leaves Serbia* at rank 2 — Pan-Slavism's degree centrality in the full 400-node graph is not overtaken by the added edge, illustrating that the injected edge is a ranking signal, not a guarantee of top rank.

B. Real-Data Louvain Validation (Partial)

As a further, non-simulated check on Step 1, we built a subgraph directly from live production data — every hyperlink target named in the running text of three collision pairs' history-volume articles (Poland, Romania, Ukraine), recorded via the same manual extraction procedure as Section III — yielding 131 nodes and 164 unique edges, a small fraction of the 403-node simulated graph of Section V-C. Running the same Louvain routine on this real subgraph (networkx, fixed seed) produced three communities of sizes 60, 44, and 27, each aligning exactly with one of the three seed articles (Poland, Romania, Ukraine) the subgraph was built from, at modularity 0.4223, despite 28 nodes being genuinely shared across two or all three articles (e.g., the Second World War, the Treaty of Versailles, NATO).

This confirms that the community-detection step executes correctly and yields a substantial, clearly non-trivial partition on real production data, not only in simulation; it covers three of the thirteen pairs and only the history-volume side, and does

not yet incorporate Steps 2–4 or the baseline comparison called for in Section V-G.

C. Search Performance

Because the production graph does not expose its complete edge list or query logs, we conducted a controlled simulation using a synthetic graph constrained by the structure reported in Section III: 403 article nodes (44 country articles, 185 from Volume 6, 174 from Volume 7), with most relationships confined within source volumes and a small number of sparse cross-volume edges. The thirteen known collisions were instantiated as eleven exact matches and two near-matches, and for the proposed method a distinct same-title, cross-volume edge was injected for each. We evaluated the degree-centrality baseline against the proposed ranking mechanism over 130 simulated queries (ten per collision pair), repeating the experiment over 30 independent random seeds and averaging the reported values across runs.

The baseline achieved an average MRR of 0.516, with Recall@1, Recall@2, Recall@5, and Recall@10 of 20.7%, 66.1%, 90.3%, and 97.8%, respectively. After introducing explicit cross-volume relationships, the proposed method achieved an MRR of 0.639 and Recall@1, Recall@2, Recall@5, and Recall@10 of 34.5%, 85.1%, 97.8%, and 99.6%, respectively.

TABLE IV. SEARCH PERFORMANCE, DEGREE-CENTRALITY BASELINE VERSUS THE PROPOSED DISAMBIGUATION-AWARE RANKING, AVERAGED OVER 30 RANDOM SEEDS (130 QUERIES PER RUN).

Metric	Baseline (degree centrality)	Proposed (cross-volume edges)
MRR	0.516	0.639
Recall@1	20.7%	34.5%
Recall@2	66.1%	85.1%
Recall@5	90.3%	97.8%
Recall@10	97.8%	99.6%

The largest relative improvement was at Recall@2 (+19.0 points), the metric most relevant to disambiguation since a successful result should surface both independently authored articles for the same entity within the top-ranked results; the gain is substantial but not perfect, consistent with the proposed edge acting as an additional ranking signal rather than a guarantee of top rank. Broken down by collision type, exact-match pairs improve from ~68% to ~88% at Recall@2, while the two near-match pairs (Portugal, Slovakia) improve from a much lower ~30% to ~52%.

The lower near-match baseline reflects the baseline's lack of any mechanism for associating non-identical titles; the proposed method improves both subsets, but the exact-match subset still outperforms after the intervention, suggesting the string-similarity threshold used in collision flagging (Section IV, Step 2) is a secondary source of error worth tuning further in production.

D. Cross-Volume Structural Coverage

Before the proposed intervention, none of the thirteen identified collision pairs had an explicit same-title relationship in the simulated graph. Following collision detection and edge injection, all thirteen pairs were connected by a dedicated cross-volume edge. Structural coverage of the known collision set therefore increased from 0% to 100% in the simulation as well, consistent with the same result already established on real data in Section V-A.

This result should not be interpreted as a 100% entity-disambiguation accuracy. Rather, it indicates that the proposed graph construction step successfully represents all known collision relationships in the evaluation set.

E. Community Detection

Louvain community detection consistently identified three major communities corresponding closely to the three source volumes. The high agreement is expected because the simulated graph preserves the strong within-volume connectivity described in the production data. Community detection is therefore treated primarily as a structural validation and community-awareness component rather than as the sole mechanism for resolving entity collisions.

TABLE V. STRUCTURAL STATISTICS OF THE SIMULATED GRAPH AND THE LOUVAIN PARTITION (SAME RUN AS TABLE 4).

Metric	Result
Nodes	403
Major communities found	3
Community purity (vs. volume label)	~99-100%
Normalized Mutual Information (NMI)	~0.99-1.00
Existing sparse cross-volume edges (pre-intervention)	8
Injected collision edges (Step 3)	13

Only eight cross-volume edges existed in the simulated graph prior to the intervention, none of which connected a known collision pair; the near-perfect purity and NMI confirm that, as anticipated in Section II-E, the pre-existing edge structure alone is a reliable proxy for volume membership, and that Louvain's role here is largely confirmatory rather than corrective. The thirteen edges injected in Step 3 are exactly the thirteen collisions catalogued in Table I, matching the 0%-to-100% coverage result reported in Section V-D.

F. Discussion

Overall, the results indicate that explicitly representing cross-volume entity relationships can improve retrieval performance, particularly when the evaluation focuses on surfacing multiple articles associated with the same entity. The improvement in Recall@2 from 66.1% to 85.1% in the simulation suggests that the injected structural signal can help overcome the absence of cross-volume links observed in the current data.

G. Remaining Validation on Production Data

The results above substitute for, but do not replace, direct measurement on the live platform. Sections III-E, V-A, and V-B make substantial progress on all three remaining steps: Section III-E now fully confirms the zero-cross-volume-link pattern for all thirteen pairs; Section V-B validates community detection on real data for three of them; and Section V-A crawls the full production text of both history volumes and the complete Countries volume to reproduce the Section V-C ranking comparison on real, rather than simulated, data, confirming the same direction of effect using a disclosed hyperlink-neighborhood proxy for retrieval rather than the platform's own live search index. Still outstanding are validating community detection (Section V-B) on all thirteen pairs rather than three, and repeating Section V-A with the platform's own search index and real user queries once accessible. We report these remaining items as ongoing work toward an extended version of this paper.

VI. Limitations

The headline ranking-improvement numbers of Section V-C remain simulation-based; Section V-A confirms the same direction of effect on a real 400-node graph and independently reproduces complete structural coverage on that graph, but using a hyperlink-neighborhood proxy for retrieval rather than the platform's own live search index, and the analysis covers a single subject area (European geography and history) out of the platform's several volume types. Full production-scale validation using the platform's live search index and real user queries is future work. Two further limitations concern scale and precision: collision detection (Step 2) was applied here to only 403 titles across three volumes, and the pairwise title comparison it requires does not by itself guarantee that every flagged pair is a genuine co-referent rather than an unrelated homonym (e.g., a person and a place sharing a name) — a risk that did not materialize among the thirteen country-name collisions studied here but must be checked when this method is extended to the platform's other, non-geographic volumes. Relatedly, the entity mentions used to motivate cross-volume linking in Section III were located by manual reading rather than automatic extraction; applying Thai named-entity recognition to the running text at scale, rather than to hyperlink anchors alone, is a natural next step and would also better align this line of work with the surrounding NLP literature.

VII. Conclusion

We have provided the first systematic empirical account of cross-corpus entity disambiguation on the Royin Encyclopedia platform, moving beyond the single anecdotal example reported previously. Working with the platform's current production data, we identified thirteen title collisions across three deployed volumes and showed, through a detailed case study, that the platform's own manually curated links never bridge these collisions even where the underlying text makes the connection explicit and the linking tool is in active use elsewhere in the same article. This result reframes cross-corpus disambiguation as a gap in the available graph structure itself, rather than a byproduct of any particular graph-construction algorithm, and motivates the domain-aware community detection approach specified in Section IV. Beyond this inventory, direct re-verification of twelve further collision pairs confirms the pattern holds throughout, and a real-data run of the proposed community-detection step (modularity 0.42, Section V-B) supports the mechanism specified in Section IV. A real 400-node graph built from the complete production text of both history volumes (Section V-A) shows this approach raising MRR from 0.885 to 0.923 and closing structural coverage of the known collision set from 0% to 100%, while making clear that the injected edge is a ranking signal rather than a guarantee of top rank; a controlled simulation (Sections V-C through V-F) grounded in the platform's reported structure confirms the same direction of effect, including the same structural-coverage result, across a much larger set of simulated queries (130 queries over 30 random seeds), with MRR rising from 0.516 to 0.639, while full validation using the platform's own live search index and real user queries remains future work.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, P10008, 2008.

[2] G. Pons, B. Bilalli, and A. Queralt, "Knowledge graphs for enhancing large language models in entity disambiguation," in The Semantic Web – ISWC 2024, Lecture Notes in Computer Science, vol. 15231, Springer, Cham, 2025.

[3] G. Lior, A. Caciularu, A. Cattan, S. Levy, O. Shapira, and G. Stanovsky, "SEAM: A stochastic benchmark for multi-document tasks," arXiv:2406.16086, 2024.

[4] F. Xie, X. Zeng, B. Zhou, and Y. Tan, "Improving knowledge graph entity alignment with graph augmentation," arXiv:2304.14585, 2023.

[5] X. Chen, T. Lu, and Z. Wang, "LLM-Align: Utilizing large language models for entity alignment in knowledge graphs," arXiv:2412.04690, 2024.

[6] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Scientific Reports, vol. 9, no. 1, 5233, 2019.

[7] Z. Dong et al., "Cross-Document Contextual Coreference Resolution in Knowledge Graphs," arXiv:2504.05767, 2025.

[8] X. Chen, Y. Cui, Z. Sun, and W. Hu, "Harnessing Structural Context for Entity Alignment Foundation Models," arXiv:2606.06109, 2026.

[9] J. Cheng, C. Lu, L. Yang, G. Chen, and F. Zhang, "EasyEA: Large language model is all you need in entity alignment between knowledge graphs," in Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20981–20995.

[10] K. Kim, Y. Jang, H. Lee, S. Kim, and Y. Ko, "Predicate importance estimation and decoupled rationale-score distillation for entity alignment," arXiv:2606.22992, 2026.

[11] J. Li et al., "A comprehensive review of community detection in graphs," arXiv:2309.11798, 2024.

[12] Z. Huang, W. Yang, X. Li, and Q. Bai, "Multi-source knowledge graph construction through LLM-assisted incremental fusion," ScienceDirect, 2026.

[13] H. Bian, "LLM-empowered knowledge graph construction: A survey," arXiv:2510.20345, 2025.

[14] V. Luantangsrisuk, K. Saengthongpattana, A. Kingkaewkanthong, K. Kriengket, and T. Supnithi, "Using knowledge graphs to enhance keyword search efficiency of the online encyclopedia and taxonomy system of the Office of the Royal Society," in Proc. 2023 18th Int. Joint Symp. Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 2023, pp. 1–4, doi: 10.1109/iSAI-NLP60301.2023.10354572.