From Travelogue to Knowledge Graph: An End-to-End Pipeline with Open-Weight LLMs for 19th-Century Dutch Travel Narratives

Tijn Donners[1], Franziska Pannach[2]

[1]University of Groningen, Netherlands
[2]Centre for Language and Cognition (CLCG), University of Groningen, Netherlands

Abstract

Travelogues offer rich historical insights but can be difficult to process using traditional NLP due to archaic language and noise. This paper presents an end-to-end pipeline using open-weight Large Language Models (LLMs) to extract, link, and model travel events based on the Academic Tourism Ontology (ATO), an extension of CIDOC-CRM. The pipeline comprises three modules: ontology-based Entity Recognition (ER) with custom tagsets, Entity Linking (EL) to Wikidata via a hybrid architecture using LLM Entity Profile Generation and strict NIL selection, and Relation and Event Extraction (RE/EE) mapping spatial relations to structured events.

Experiments across eight models show Dutch few-shot prompts yield the highest ER performance (kimi-k2.7-code F1=0.896), while gemma4:31b offers the best speed-accuracy trade-off. For EL, gemma4:31b achieves strong Wikidata linking (F1=0.764) and NIL prediction (F1=0.84). In RE/EE, LLMs excel at binary visited/mentioned classification (F1=0.924) but struggle to map entities into actual event structures (F1=0.785) and fine-grained relations (F1=0.686). To model proximate locations, a new ATO property, *happened_in_proximity_of*, is introduced. Overall, open-weight LLMs can transform historical travelogues into ATO-compliant knowledge graphs, offering a reproducible, resource-efficient blueprint for cultural heritage, though human validation of nuanced relations remains advisable.

Keywords

Large Language Models, Entity Recognition, Entity Linking, Relation Extraction, Knowledge Graphs, Long-Tail Entities, CIDOC-CRM

CEUR-WS.org/Vol-4340/paper20.pdf

1. Introduction

In 1809, the Dutch nobleman *Theodorus van Swinderen* travelled through Germany with fellow students, documenting their journey in a published travelogue. This work functioned as a blueprint for future travellers, such as *Reneke de Marees van Swinderen* in 1816 and *Ulrich Herman Wielinga Huber* in 1819. These 19th-century epistolary travelogues constitute a fruitful corpus offering historians insights into intellectual and cultural self-cultivation, historical mobility, and cross-cultural encounters during a tumultuous Napoleonic era. They are characterised by a high density of long-tail entities: uncommon locations, specific museum artefacts, minor physical structures, and localised historical figures that are underrepresented in modern knowledge bases and language model training data.

Automatically extracting those entities presents multiple challenges. Historical texts feature archaic spelling variations (e.g., 'Ocksendorf' and 'Ossendorf'), OCR/HTR noise, complex narrative structures, and long-tail entities that are underrepresented in the training data of modern Natural Language Processing (NLP) models. To make these narratives semantically machine-readable and interoperable, they should be mapped to formal ontologies like the CIDOC Conceptual Reference Model (CIDOC-CRM) and its domain-specific extension, the Academic Tourism Ontology (ATO).[1] The rise of Large Language Models (LLMs) has prompted some to question whether the extensive work of transforming unstructured text into Knowledge Graphs (KGs) is still worthwhile [1]. However, while LLMs possess

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters)
contact@tijndonners.nl (T. Donners); f.a.pannach@rug.nl (F. Pannach)
0009-0006-6439-5250 (T. Donners); 0000-0003-4216-8410 (F. Pannach)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://github.com/LuotongCheng/The_Academic_Tourism_Ontology_ATO/blob/main/ATO.rdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

vast parametric knowledge, they exhibit hallucinations and a diminished capacity to retrieve long-tail knowledge. Recent benchmarking by Sun et al. [2] demonstrates that GPT-4 achieved an average accuracy of only 31% on factual queries, with a sharp decline in performance as entity popularity decreases. Because historical domains are abundant in less-popular, long-tail entities, LLMs cannot and should not replace KGs. It is essential to preserve factual long-tail knowledge in KGs, as this knowledge is suppressed in the parametric memory of LLMs due to inherent popularity bias [3, 4]. Additionally, because KGs are updatable, LLMs do not have to be retrained to update their parametric knowledge; they can access updated knowledge via KGs [5]. By mapping unstructured text into a KG, researchers can leverage the linguistic strengths of LLMs for extraction while maintaining a structured, verifiable source of truth that preserves factual long-tail knowledge.

This research builds and evaluates an end-to-end NLP pipeline integrating open-weight LLMs for Entity Recognition (ER), Entity Linking (EL), and Relation/Event Extraction (RE/EE), modelling the data as linked data in the Resource Description Framework (RDF) aligned with the ATO, specifically focusing on toponyms.[2] To mitigate the popularity bias and long-tail retrieval limitations of LLMs, the pipeline employs a hybrid EL architecture combining LLM-based Entity Profile Generation (EPG) [6] and a strict selector that abstains from linking unresolvable entities. This study addresses the research gap of applying LLM prompting to long-tail entities in historical Dutch texts by asking: How effectively can an end-to-end NLP pipeline utilising open-weight LLMs extract, link, and model long-tail entities and travel events from 19th-century Dutch travelogues into ATO-based KGs? To answer this, the study evaluates the trade-offs between computational cost, inference time, and extraction accuracy across eight open-weight models, demonstrating that LLM-driven pipelines can lower the barrier to unlocking large collections of historical documents while preserving data sovereignty for cultural heritage institutions.

2. Related Works

Named Entity Recognition (NER) on historical texts faces domain heterogeneity, noise, and low resources [7] . While fine-tuning Dutch BERT models yields high performance on specific historical corpora [8, 9, 10], coarse-grained tagsets and data scarcity limit their applicability to 19th-century Dutch travelogues. Alternatively, LLMs using zero-/few-shot prompting achieve strong results on historical German and French NER, excelling at 'non-rigid designators' [11, 12]. Non-LLM zero-shot alternatives like GLiNER [13, 14] support custom tagsets but lack multilingual schema support. Because our pipeline requires capturing entities, among which are non-rigid designators, to map to a fine-grained ontology, we evaluate LLMs for entity extraction beyond traditional NER.

EL bridges historical entities to knowledge bases but suffers from OCR/HTR noise and poor long-tail entity coverage [3, 4]. Standard linkers fail on historical datasets, whereas LLM zero-shot prompting proves more robust. Our corpus exhibits a severe "long tail" of unpopular, underrepresented entities. Traditional EL pipelines separate candidate generation and ranking [15], with zero-shot BERT rerankers aiding low-resource linking [16] . Recently, LLMs have been integrated to normalize ambiguous strings before retrieval [17] via EPG [6]. EPG normalizes archaic spellings and enables complex querying by contextualizing the entity before candidate generation, maximizing long-tail recall. However, EPG risks LLM popularity bias for long-tail entities. To mitigate this, frameworks like LELA and OneNet use LLMs as strict selectors with NIL assignment capabilities to ensure precision [18, 19]. This research proposes a hybrid approach: EPG for broad candidate generation, followed by LLM-based reranking and strict selection.

RE and EE map unstructured text to Semantic Web formats [20, 21] using ontologies like CIDOC-CRM [22, 23]. For travelogues, the ATO extends CIDOC-CRM with domain-specific travel events. A major RE challenge is distinguishing visited locations from mere mentions, classifying them into fine-grained relations (IN, NEAR, THRU, TO, FROM) [24]. Prior work on modern texts using simpler FROM/TO relations [25] or implicit/explicit presence evidence scales [26] informs our approach. Unlike standard

[2] https://github.com/tijn-donners/hist-dutch-travelogues-nlp

pipeline or joint learning RE predicting simple triples [27], ATO extraction requires anchoring multiple entities to a central conceptual travel event. This complexity, alongside fluid boundaries between spatial classes, complicates traditional triple-based F1 evaluation, necessitating both binary visited/mentioned and fine-grained relational assessments based on the generated Knowledge Graph.

3. Data

The pipeline accommodates three 19th-century Dutch epistolary travelogues at varying digitization stages (1809 printed, 1816 transcribed, 1819 handwritten). The 1816 travelogue by Reneke de Marees van Swinderen serves as the primary evaluation dataset due to its high-quality manual transcription. Its third letter (315 entities) acts as the main test set. A custom ontology-based tagset covers core travel entities (see Table 1). Entities were manually annotated in Recogito and exported as JSONLD.

Table 1
Custom CIDOC-CRM and ATO based tagset used for the Entity Recognition module.

Label	Description	Example Span (translated)
E53_Place	Abstract locations and spatial extents (GPEs)	“Kassel”, “the second room”
E18_Physical_Thing	Tangible physical buildings and structures	“Wilhelmshöhe”
Mode_of_Transportation	Transport modes as verbs (walking) or nouns (carriage)	“walked”, “carriage”
E52_Time-Span	Temporal expressions	“the next morning”, “5 hours”
E19_Physical_Object	Physical, movable objects	“painting”
E20_Biological_Object	Movable objects consisting of biological material	“taxidermied birds”
F2_Expression	Immaterial, conceptual expressions	“third letter”
E31_Document	Specific, identifiable documents	“Edict of Nantes”

Few-shot examples were derived from the 4th and 5th letters based on a model-agnostic zero-shot error analysis. This targeted approach addresses frequent LLM misclassifications, such as confusing spatial extents with physical buildings, or conceptual expressions with specific historical documents.

For EL, 170 location entities were manually linked to Wikidata and GeoNames. Highlighting the long-tail challenge, 45.3% of these locations lacked KB entries and were assigned NIL. The popularity distribution of the other 53.7% of manually linked location entities can be seen in Figure 1, where the y-axis shows the entity popularity based on pageviews of Wikidata entities across Wikimedia platforms, and x-axis show the popularity ranking of the same entities.

For RE/EE, each location was assigned a fine-grained person-location relation (IN, NEAR, THRU, TO, FROM or NO_REL) [24]. RE/EE few-shot examples were generated via deepseek-v4-pro and manually verified to prevent data leakage in the test set.

4. Method

4.1. Entity Recognition

The ER module extracts ontology-typed entities, including non-named entities like ’wagen’ (Dutch for ’carriage’), to map directly to the ATO, in this case *Mode_of_Transportation*. Standard NER models lack custom tagset support, so the multilingual BERT model GLiNER was used to establish a baseline, yielding a relaxed F1 of 0.3.

To surpass this baseline, an LLM-based ER module was implemented using spaCy, spacy-llm, and Ollama (LLM backend) for reproducible local or cloud inference. Eight open-weight LLMs—gemma4:31b

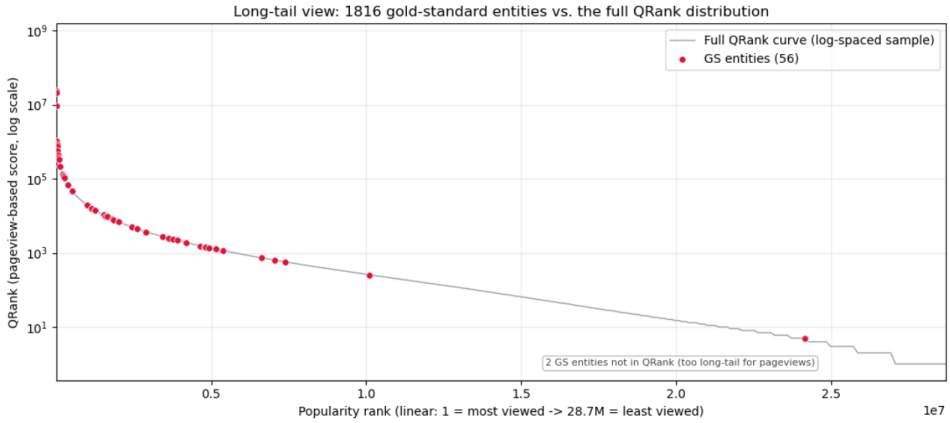


Figure 1: Long-tail view of the EL evaluation data. Popularity rank on the x-axis (linear) and a pageview-based QRank on the y-axis (logarithmic). The grey line represents the QRank popularity curve. The red dots indicate the gold standard annotations' position on the QRank popularity curve.

(31B), kimi-k2.7-code (1T), qwen3.5:397b (397B), cogito-2.1:671b (671B), mistral-large-3:675b (675B), glm-5.1 (754B), deepseek-v4-flash (284B), and deepseek-v4-pro (1.6T)—were evaluated across 96 configurations. These configurations varied temperature (0.0, 0.5, 1.0) [28], prompt language (Dutch vs. English), and shot count (zero-shot vs. few-shot). Inference was run on datacenter grade hardware, using Ollama cloud API. The code is theoretically configurable to run locally, but most models used in this research will not run (quickly) on consumer grade hardware.

Prompts constructed via Jinja templates use prompt engineering strategies [11] by initializing a historian persona and enforcing a strict pipe-delimited output schema. Few-shot examples were derived from expert annotations of the 4th and 5th letters. A limitation is the reliance on a single annotator for the evaluation set, precluding inter-annotator agreement metrics.

The pipeline processes text page-by-page to limit context window size and uses a retry mechanism for API resilience. Outputs are serialized into .spacy files for downstream tasks. ER performance is evaluated via strict and relaxed precision, recall, and F1-scores against the gold standard, with detailed CSVs generated for error analysis.

4.2. Entity Linking

The EL module links toponyms (E53_Place and E18_Physical_Thing) to Wikidata and GeoNames. Time-spans and transportation modes are too generic for disambiguation. Following a LELA-inspired architecture combined with EPG [18, 6], the pipeline consists of three LLM submodules: candidate generation, pointwise reranking, and candidate selection. The same LLM is used across all stages to evaluate end-to-end performance. Seven open-weight models were tested (gemma4:31b, kimi-k2.7-code, qwen3.5:397b, mistral-large-3:675b, glm-5.1, deepseek-v4-flash, deepseek-v4-pro), with additional thinking-mode experiments (low, medium, high) on the top three performing models (deepseek-v4-flash, gemma4:31b, kimi-k2.7-code).

The candidate generation submodule uses a Dutch prompt to generate an EPG profile containing a modernized German name, region/country, and descriptive keywords. This enables a dual-path retrieval strategy: a sparse lane for direct string matching via the wbsearchentities API, and a dense lane utilizing the EPG profile for a Wikidata SPARQL query, an English Wikipedia full-text search, and a local fuzzy-matched German GeoNames gazetteer.

The reranking submodule prompts the LLM with the entity surface form and page context to assign a pointwise rank (0-10) to candidates, passing the top three to the selection stage. To address the long-tail problem, the selection LLM is explicitly prompted to return NIL if the correct candidate is absent, mitigating false positives. Final candidates are enriched with Wikipedia extracts or mapped GeoNames feature descriptions to provide context for selection.

Performance is evaluated via precision, recall, and F1 for Wikidata predictions. GeoNames is excluded from evaluation, serving only as a coordinate fallback. The pipeline outputs a serialized .spacy file with linked identifiers for downstream RE.

4.3. Relation and Event Extraction

The RE/EE module transforms ER and EL outputs into structured KGs serialized in Turtle RDF. The graph is governed by the ATO, an event-centric extension of CIDOC-CRM [29].

Extraction is performed by LLMs prompted in Dutch with the annotated text, ontology specifications, and a strict JSON output schema. Jinja templates provide seven few-shot examples. The JSON output is parsed and transformed into RDF using rdflib. Models are served via Ollama at temperature 0.0, matching those used in the EL stage with the addition of glm-5.2.

Evaluation uses a two-tiered strategy against a gold standard recording manually annotated categories. Tier 1 is a binary F1 score distinguishing visited from non-visited locations, computed from both LLM JSON and generated RDF to validate graph structural integrity. Tier 2 is a per-label F1 analysis for the six relation categories.

The ATO models journeys as four event types mapped to CIDOC-CRM: translocations (*E9_Move*) using *P27_moved_from* and *P26_moved_to*, and indoor tours, outdoor tours, and stays (*E7_Activity*) using *P7_took_place_at* and *P8_took_place_on_or_within*. Events are ordered temporally (*P183_ends_before_the_start_of*) and mereologically (*P10_contains*) to form a letter-level itinerary. The Tour Group is modeled as a static *E74_Group* participant.

Following Soni et al. [24], locations are assigned to six spatial categories: IN, NEAR, THRU, TO, FROM, or NO_REL. Most map to standard CIDOC-CRM properties (e.g., TO to *P26_moved_to*). THRU uses nested sub-moves, and NO_REL links only via *P67_refers_to*. Because CIDOC-CRM lacks a proximity property, the custom *happened_in_proximity_of* property was added to the ATO for the NEAR category.

5. Results

Before reporting the results, it is essential to clarify what evaluation metrics are used. Model performance on the NER, EL, and RE tasks is evaluated using precision, recall, and F1-score, calculated from the confusion matrix of true positives (TP, correct predictions by the model), false positives (FP, the model assigns a label while it shouldn't have), true negatives (TN, the model correctly restrained from assigning a label), and false negatives (FN, missed predictions) [30]. Precision measures the degree of correctness in model predictions by penalising over-generation and measuring the proportion of false positives (TP / (TP + FP)). Recall measures completeness of predictions by calculating what proportion of the gold-standard entities are actually identified by the prediction model (TP / (TP + FN)). Put more plainly, more false negative predictions result in a lower recall (completeness), while more false positive predictions lower precision (correctness). The F1-score, defined as the harmonic mean of precision and recall rather than the average, is reported as a single summary statistic that balances these two [30]. This makes F1 particularly well-suited to compare models across the three tasks, since it reflects both the completeness and the correctness of extracted information. The F1 score reported here uses a relaxed span-matching approach. In this method, a predicted span is considered a true positive if it overlaps with the gold-standard span, even if it doesn't match exactly. To give a practical example, if the model predicts the entity type *Mode_of_Transportation* for the span "the large carriage" while the gold-standard entity span is "carriage", this is still considered a true positive prediction. This relaxed evaluation is particularly useful in historical domains, where the precise inclusion or exclusion of articles or adjectives is less critical than correctly identifying the core entity. Additionally, an ontology-based tagset may also introduce more flexible entity boundaries, further justifying the use of relaxed F1 scoring.

5.1. Entity Recognition

All eight evaluated LLMs outperformed the multilingual BERT baseline (GLiNER, F1=0.3). Across the 96 experimental configurations, few-shot prompting consistently outperformed zero-shot baselines, yielding an average F1 increase of +0.073. This validates the strategic inclusion of few-shot examples specifically designed to target common model-agnostic errors, such as confusing physical buildings with spatial extents. Furthermore, Dutch-language prompts enhanced extraction quality over English prompts in the majority of configurations (35 out of 48). Conversely, temperature settings exerted negligible influence on mean F1 scores. To maximize reproducibility, a temperature setting of 0.0 is recommended.

The optimal configuration consists of Dutch few-shot prompts at 0.0 temperature. Under these settings, kimi-k2.7-code achieved the highest score (F1=0.896), followed closely by gemma4:31b (F1=0.875). While kimi-k2.7-code offers marginally superior accuracy, gemma4:31b achieves competitive results with lower computational cost and faster inference time, making it the pragmatic choice for large-scale processing of historical corpora.

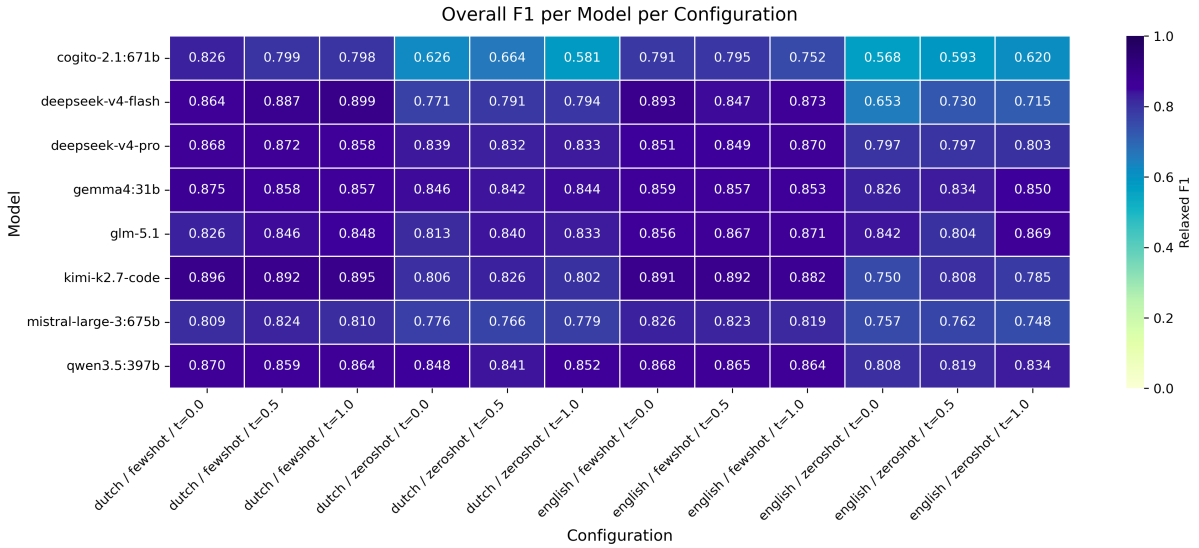


Figure 2: Heatmap of relaxed F1 scores as values, LLMs on the y-axis and configurations on the x-axis.

Entity-level analysis confirms that the travel-relevant labels fundamental to the ATO are recognized with high reliability. *E53_Place*, *E52_Time-Span*, *E18_Physical_Thing*, and *Mode_of_Transportation* all achieved F1 scores above 0.75 across all models, indicating that geographical, temporal, and transport references are relatively unambiguous. In contrast, the abstract and sparsely annotated labels *E31_Document* and *F2_Expression* remained challenging, yielding low and variable scores. Within the *E18* to *E19* to *E20* taxonomic hierarchy, models occasionally defaulted to the general superclass *E18* rather than the specific subclass *E19*. However, both recommended models performed well across this hierarchy, demonstrating robust disambiguation of ontologically nested entity types. These results confirm that optimized LLM-based ER is effective for 19th-century Dutch travelogues.

5.2. Entity Linking

The EL architecture was evaluated on 170 toponyms (*E53_Place* and *E18_Physical_Thing*) from the 1816 travelogue. Evaluation measured end-to-end F1 across the three pipeline stages and assessed NIL prediction to address the long-tail entity problem [3, 4]. Because the pipeline requires three LLM calls per entity, thinking mode was disabled to manage computational costs. In non-thinking mode, gemma4:31b achieved the highest F1 score (0.764) and demonstrated exceptional NIL prediction (F1=0.84). Conversely, mistral-large-3 was excluded due to unparsable outputs in non-thinking mode. While enabling thinking

mode on models like deepseek-v4-flash yielded a +0.13 F1 improvement, it increased inference time from 20 minutes to 2 hours. Consequently, gemma4:31b in non-thinking mode was selected as the recommended configuration.

Table 2
Performance metrics per LLM for EL to Wikidata. Divided between EL performance and NIL prediction performance.

	Wikidata EL			Wikidata NIL		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
deepseek-v4-flash	0.5826	0.7204	0.6442	0.855	0.610	0.712
deepseek-v4-pro	0.5950	0.7742	0.6729	0.816	0.520	0.635
gemma4:31b	**0.8000**	0.7312	**0.7640**	0.800	**0.883**	**0.840**
glm-5.1	0.5725	**0.8065**	0.6696	**0.897**	0.455	0.603
kimi-k2.7-code	0.5271	0.7312	0.6126	0.878	0.468	0.610
mistral-large-3:675b	0.5752	0.6989	0.6311	0.719	0.533	0.612
qwen3.5:397b	0.6989	0.6989	0.6989	0.766	0.766	0.766

While gemma4:31b excels at precision and correctly assigning NIL values to long-tail location entities, larger models like glm-5.1 and deepseek-v4-pro achieve higher recall. This higher recall is likely caused by their superior performance during the candidate generation stage. As substantially larger models, they may possess broader internal knowledge representations, allowing them to generate more accurate Entity Profiles for ambiguous or archaic entities, which in turn retrieves a more complete pool of candidates. This contrast suggests a future hybrid approach where a high-recall model handles EPG, and gemma4:31b handles reranking and selection.

Table 3 demonstrates a discrepancy in EL performance between Head (F1 = 0.93) and Tail entities (F1 = 0.71). Using a QRank threshold of 1 million places the 1 million most popular entities into the Head category (N = 47), leaving the remaining entities in the Tail category (N = 46). Entities with a gold-standard NIL label are excluded, as they cannot be categorised into either group.

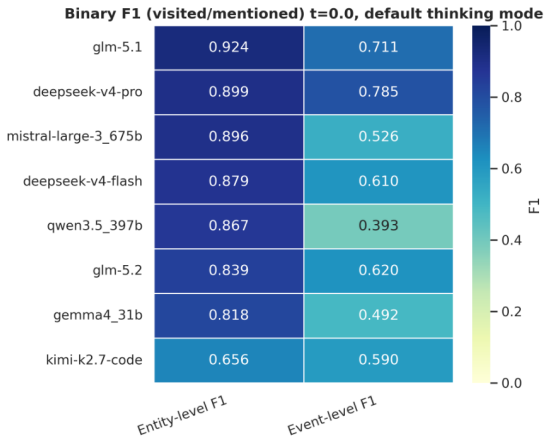
Table 3
Performance comparison of various language models across Head entities (QRank < 1e6) and Tail entities (QRank > 1e6) and all entities. Entities with NIL predictions in the gold standard data are left out in the Head and Tail metrics.

	Head Evaluation ($N = 47$)			Tail Evaluation ($N = 46$)		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
deepseek-v4-flash	0.9348	**0.9149**	0.9247	0.6154	0.5217	0.5647
deepseek-v4-pro	**0.9556**	**0.9149**	**0.9348**	0.7436	0.6304	0.6824
gemma4:31b	**0.9556**	**0.9149**	**0.9348**	**0.8065**	0.5435	0.6494
glm-5.1	**0.9556**	**0.9149**	**0.9348**	0.7273	**0.6957**	**0.7111**
kimi-k2.7-code	0.9130	0.8936	0.9032	0.6190	0.5652	0.5909
mistral-large-3:675b	0.9318	0.8723	0.9011	0.7273	0.5217	0.6076
qwen3.5:397b	0.9268	0.8085	0.8636	0.7941	0.5870	0.6750

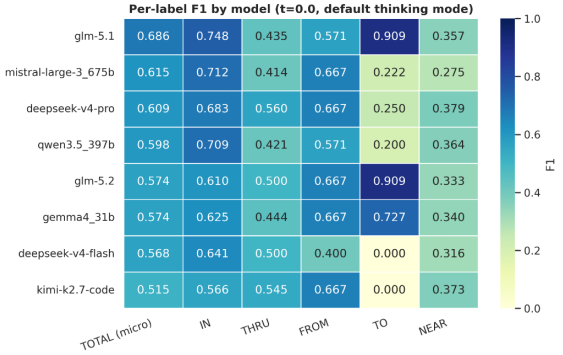
End-to-end evaluation combining the recommended ER model (kimi-k2.7-code) and gemma4:31b for EL yielded an F1 score of 0.7 (P = 0.72, R = 0.67). Error propagation from the ER stage naturally caused this decrease, providing a realistic performance metric for a fully automated pipeline without human intervention. Ultimately, gemma4:31b is preferred for its optimal balance of accuracy, superior NIL performance, and computational efficiency. However, gemma4:31b was significantly outperformed in terms of recall for the Tail entities by glm-5.1, confirming the need for hybrid approach with multiple models, utilizing the precision and NIL assignment capabilities of gemma4:31b and the parametric

knowledge of larger Mixture of Experts (MoE) models for long-tail recall.

5.3. Relation and Event Extraction



(a) Heatmap of binary (visited/non-visited) F1 per model for entity-level relation extraction and event-level (RDF derived) relation extraction.



(b) Heatmap of entity-level fine-grained F1 per relation type (x-axis) and LLM (y-axis) including the micro average F1 per model. Note that THRU, FROM and TO have insufficient gold-standard annotations.

Figure 3: Comparative analysis of Relation Extraction performance.

A clear performance gap emerges between entity-level and event-level binary classification. LLMs effectively identify visited entities at the entity level (glm-5.1 reaching 0.924 F1), but struggle when linking these entities to travel events in the generated RDF graph (deepseek-v4-pro leading with an F1 of 0.785). This disparity stems from the models' abstention from linking all extracted locations and relations to the Knowledge Graph. While entity-level evaluation assesses the raw initial extractions, event-level metrics strictly evaluate the entities extracted from generated KG.

Performance drops significantly for fine-grained relation classes (IN, NEAR, FROM, TO, THRU, NO_REL), with the best model (glm-5.1) reaching an F1 of 0.686. This decline is largely attributable to the inherent ambiguity of spatial relations. Error analysis reveals the most frequent misclassification involves NEAR locations predicted as IN, highlighting the difficulty of distinguishing proximity from actual presence in historical travelogues. Furthermore, the low frequency of THRU, TO, and FROM labels in the gold standard precludes reliable evaluation for these categories.

In conclusion, while LLMs demonstrate clear capability for binary visited/non-visited extraction from historical travelogues, they struggle to reliably map these entities into coherent RDF travel events. Fine-grained relation classification remains a significant challenge requiring more extensive gold-standard data with calculated inter-annotator agreement to differentiate model errors from genuine interpretive ambiguity.

6. Conclusion

This research developed and evaluated an NLP framework assessing open-weight LLMs for ontology-based ER, long-tail EL, and RE/EE from 19th-century Dutch travelogues. The framework maps extracted data into the ATO, an extension of CIDOC-CRM. A primary motivation for this study was addressing the persistent challenge of long-tail entities in historical NLP, where archaic spellings, obscure locations, and sparse knowledge base coverage limit traditional extraction pipelines.

Evaluating 96 configurations across eight LLMs revealed that Dutch few-shot prompts, designed via zero-shot error analysis, consistently yielded the best ER results. kimi-k2.7-code achieved the highest F1 (0.896), while gemma4:31b delivered competitive performance (F1=0.875) at a fraction of the computational cost. Performance on downstream-relevant labels (*E53_Place*, *E18_Physical_Thing*,

E52_Time-Span, and *Mode_of_Transportation*) was strong. The models also successfully distinguished taxonomic lineages (*E18* to *E19* to *E20*), though conceptually ambiguous classes like *F2_Expression* and *E31_Document* remained challenging. This indicates that while LLMs possess robust semantic discrimination for clear-cut ontological boundaries, abstract classes lacking sufficient training examples remain difficult to extract reliably from historical text.

The EL module targeted the long-tail problem by integrating EPG to increase recall for lesser-known entities, paired with a strict selector LLM to assign NIL values, thereby mitigating false positives. The hybrid architecture deployed EPG to contextualize and modernize entity strings, unearthing candidates that would otherwise remain undiscovered, while the selector LLM enforced strict abstention to prevent hallucinated links. gemma4:31b achieved strong Wikidata linking (F1=0.764) and exceptional NIL prediction (F1=0.84) in its default non-thinking mode, demonstrating a conservative linking policy crucial for data integrity. Conversely, deepseek-v4-flash reached a slightly higher F1 of 0.774 when utilizing high thinking mode, though this incurred a substantial computational cost. This research also confirmed the expected: a substantial performance gap between popular (F1 = 0.93) and long-tail entities (F1 = 0.71).

While this architecture performs well for geographical toponyms, its evaluation is limited to this entity type. Critically, combining EPG for broad candidate generation with a strict LLM selector suggests an adaptable framework. This hybrid EL architecture should be tested on other entity types, such as persons and artifacts, to evaluate to what extent this methodology generalises across entity types and domains.

For RE/EE, models assigned fine-grained spatial relations (IN, NEAR, THRU, FROM, TO, NO_REL) to toponyms to anchor them to structured travel events. A custom ATO property, *happened_in_proximity_of*, was introduced for the NEAR category. LLMs excelled at binary visited versus non-visited classification (glm-5.1 F1=0.924) but exhibited discrepancy when mapping these entities to actual travel events in the RDF (deepseek-v4-pro event-level F1=0.785). Furthermore, fine-grained classification proved difficult, particularly distinguishing NEAR from IN and NO_REL. The inherent interpretive ambiguity of historical travel narratives means that a single location can be modeled as both a point of arrival and a point of departure. Future iterations should separate RE and EE into two distinct LLM calls to simplify the complex prompt, thereby improving instruction following and event mapping. Compared to the ER and EL modules, RE/EE and KG generation remains the most challenging part of this pipeline.

Overall, the framework demonstrates that open-weight LLMs are viable for modeling historical travel events as linked data, though a human-in-the-loop remains necessary to validate semantic precision. The primary limitations include reliance on a single-annotator gold standard, small test sets, a temperature configuration bug during EL thinking mode experiments, and the static modeling of travel groups. Future work involves creating a multi-annotator evaluation set, conducting EL ablation studies to isolate the impact of EPG and the selector LLM on long-tail retrieval, and integrating the *happened_in_proximity_of* property into the ATO. Ultimately, this pipeline offers a reproducible alternative to proprietary models, lowering the barrier to unlocking large cultural heritage collections and transforming unstructured historical texts into queryable KGs.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini and Ollama in order to: Grammar and spelling check. Further, the authors used Gemini and Ollama for: Drafting content, Paraphrase and reword, and Improve writing style. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, D. Graux, Large language models and knowledge graphs: Opportunities and challenges, Transactions on Graph Data and Knowledge (TGDK) 1 (2023) 2:1–2:38. doi:10.4230/TGDK.1.1.2.

[2] K. Sun, Y. Xu, H. Zha, Y. Liu, X. L. Dong, Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs?, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 311–325. doi:10.18653/v1/2024.naacl-long.18.

[3] A. Graciotti, L. Piano, N. Lazzari, E. Daga, R. Tripodi, V. Presutti, L. Pompianu, KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 20316–20339. doi:10.18653/v1/2025.findings-acl.1042.

[4] A. Graciotti, N. Lazzari, V. Presutti, R. Tripodi, Musical heritage historical entity linking, Artificial Intelligence Review 58 (2025) 140. doi:10.1007/s10462-024-11102-9.

[5] E. Lavrinovics, R. Biswas, J. Bjerva, K. Hose, Knowledge graphs, large language models, and hallucinations: An NLP perspective, Journal of Web Semantics 85 (2025) 100844. doi:10.1016/j.websem.2024.100844.

[6] T. Lai, H. Ji, C. Zhai, Improving candidate retrieval with entity profile generation for Wikidata entity linking, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, 2022, pp. 3696–3711. doi:10.18653/v1/2022.findings-acl.292.

[7] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification in historical documents: A survey, ACM Comput. Surv. 56 (2023). doi:10.1145/3604931.

[8] V. Provatorova, M. van Erp, E. Kanoulas, Too young to NER: Improving entity recognition on dutch historical documents, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, ELRA and ICCL, 2024, pp. 30–35. URL: https://aclanthology.org/2024.lt4hala-1.4/.

[9] S. I. Arnoult, L. Petram, P. Vossen, Batavia asked for advice: Pretrained language models for named entity recognition in historical texts, in: S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, S. Szpakowicz (Eds.), Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, 2021, pp. 21–30. doi:10.18653/v1/2021.latechclfl-1.3.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[11] T. Hiltmann, M. Dröge, N. Dresselhaus, T. Grallert, M. Althage, P. Bayer, S. Eckenstaler, K. Mendi, J. M. Schmitz, P. Schneider, W. Sczeponik, A. Skibba, NER4all or context is all you need: Using LLMs for low-effort, high-performance NER on historical texts. a humanities informed approach, arXiv preprint (2025). doi:10.48550/ARXIV.2502.04351. arXiv:2502.04351.

[12] S. Galka, G. Vogeler, Annotating, projecting, and interpreting named entities in digital scholarly editions with LLMs, International Journal of Digital Humanities 7 (2026) 483–509. doi:10.1007/s42803-025-00114-8.

[13] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney, A. Lewis, GLiNER2: Schema-driven multi-task learning for structured information extraction, in: I. Habernal, P. Schulam, J. Tiedemann (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2025, pp. 130–140. doi:10.

18653/v1/2025.emnlp-demos.10.

[14] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 5364–5376. doi:10.18653/v1/2024.naacl-long.300.

[15] F. Shanaz, R. G. Ragel, Wikidata based location entity linking, in: Proceedings of the 2020 9th International Conference on Software and Computer Applications, 2020, pp. 307–312. doi:10.1145/3384544.3384592.

[16] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 6397–6407. doi:10.18653/v1/2020.emnlp-main.519.

[17] D. Vollmers, H. M. Zahera, D. Moussallem, A.-C. N. Ngomo, Contextual augmentation for entity linking using large language models, 2025. URL: https://arxiv.org/abs/2510.18888. arXiv:2510.18888.

[18] S. Haffoudhi, F. M. Suchanek, N. Holzenberger, LELA: An LLM-based entity linking approach with zero-shot domain adaptation, arXiv preprint (2026). doi:10.48550/arXiv.2601.05192. arXiv:2601.05192.

[19] X. Liu, Y. Liu, K. Zhang, K. Wang, Q. Liu, E. Chen, OneNet: A fine-tuning free framework for few-shot entity linking via large language model prompting, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13634–13651. URL: https://aclanthology.org/2024.emnlp-main.756/. doi:10.18653/v1/2024.emnlp-main.756.

[20] C. Bizer, T. Heath, T. Berners-Lee, Linked Data - The Story So Far, 1 ed., Association for Computing Machinery, New York, NY, USA, 2023, p. 115–143. URL: https://doi.org/10.1145/3591366.3591378.

[21] E. Hyvönen, Cultural Heritage on the Semantic Web, Springer International Publishing, Cham, 2012, pp. 1–11. URL: https://doi.org/10.1007/978-3-031-79438-4_1. doi:10.1007/978-3-031-79438-4_1.

[22] G. Bruseker, N. Carboni, A. Guillem, Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM, Springer International Publishing, Cham, 2017, pp. 93–131. URL: https://doi.org/10.1007/978-3-319-65370-9_6. doi:10.1007/978-3-319-65370-9_6.

[23] M. Doerr, The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92. doi:10.1609/aimag.v24i3.1720.

[24] S. Soni, A. Sihra, E. Evans, M. Wilkens, D. Bamman, Grounding characters and places in narrative text, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11723–11736. URL: https://aclanthology.org/2023.acl-long.655/. doi:10.18653/v1/2023.acl-long.655.

[25] D. Zeginis, E. Kalampokis, K. Tarabanis, Applying an ontology-aware zero-shot llm prompting approach for information extraction in greek: the case of diavgeia gov gr, in: Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics, PCI '24, Association for Computing Machinery, New York, NY, USA, 2025, p. 324–330. URL: https://doi.org/10.1145/3716554.3716603. doi:10.1145/3716554.3716603.

[26] J. Opitz, C. Raclé, E. Boros, A. Michail, M. Romanello, M. Ehrmann, S. Clematide, Clef hipe-2026: Evaluating accurate and efficient person–place relation extraction from multilingual historical texts, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 354–363.

[27] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, ACM Comput. Surv. 56 (2024). URL: https://doi.org/10.1145/3674501. doi:10.1145/3674501.

[28] Y. Song, G. Wang, S. Li, B. Y. Lin, The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 4195–4206. URL: https://aclanthology.org/2025.naacl-long.211/. doi:10.18653/v1/2025.naacl-long.211.

[29] R. Studer, V. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & Knowledge Engineering 25 (1998) 161–197. URL: https://www.sciencedirect.com/science/article/pii/S0169023X97000566. doi:https://doi.org/10.1016/S0169-023X(97)00056-6.

[30] D. M. W. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, ArXiv abs/2010.16061 (2011). URL: https://api.semanticscholar.org/CorpusID:3770261.