Semantic Enrichment of Historical Texts for Cultural Tourism Using Named Entity Recognition, Ontologies, and Large Language Models

Fabio Clarizia[1], Massimo De Santo[2], Rosario Gaeta[3], Rocco Loffredo[4]

***Abstract*—This paper presents a novel framework for the semantic enrichment of historical texts in the context of cultural tourism. We integrate Named Entity Recognition (NER), domain-specific ontologies, and prompt-engineered Large Language Models (LLMs) to enable contextual understanding, disambiguation, and inferential reasoning over unstructured heritage documents. The system supports interactive exploration of annotated texts, revealing hidden connections between places, historical figures, events, and cultural practices across time. We discuss current limitations, such as ontology incompleteness and LLM hallucinations, and outline future work including multilingual expansion, crowd-sourced ontology enrichment, and mobile augmented reality deployment.**

***Index Terms*—Semantic annotation; Named Entity Recognition; Large Language Models; Ontologies; Cultural tourism; Digital humanities; Knowledge graphs; Heritage informatics; Historical text analysis; Smart tourism systems**

I. Introduction

The fusion of digital technologies and humanities has led to groundbreaking advancements in the analysis and interpretation of historical data. One of the most promising areas of interdisciplinary research is the application of Natural Language Processing (NLP) techniques, particularly Named Entity Recognition (NER), ontologies, and Large Language Models (LLMs), to the semantic annotation and contextual understanding of ancient texts. This integration is proving particularly valuable in the field of cultural tourism, where the ability to extract, link, and enrich historical information can offer novel experiences to both researchers and tourists. Tourism, especially cultural and heritage tourism, is undergoing a digital transformation. Visitors today demand more than just passive observation, they seek immersive, meaningful narratives that connect historical sites to broader cultural and historical contexts [1]. However, the vast majority of the world's historical heritage is documented in unstructured formats: manuscripts, travel logs, inscriptions, epigraphs, or early guidebooks that lack formal annotation or semantic links. These documents often contain ambiguities, archaic language, and historical references that are challenging to interpret without expert knowledge. NER allows us to identify people, places, organizations, dates, and artifacts within these texts [2], [3], acting as a first step toward structured understanding. However, NER alone is not sufficient. Many entities have multiple meanings depending on context. To disambiguate these entities and relate them to structured knowledge, ontologies play a crucial role. Ontologies model concepts, relationships, and attributes within a domain, enabling machines to link "Alexandria" to the correct historical period, geography, or associated individuals [4]–[6]. LLMs such as GPT-4, BERT, and their fine-tuned derivatives bring a new dimension: the ability to infer unstated relationships and fill in contextual gaps. For example, given a Roman text referencing a "praetor" visiting a temple, an LLM, guided by an appropriate ontology, can infer the temporal range, political authority, and even social customs related to that visit. This inferential capacity is crucial when dealing with incomplete, ambiguous, or poetic historical texts [7], [8]. In this paper, we argue that the combination of these three technologies, NER, ontologies, and LLMs can radically enhance how historical texts are analyzed and presented, particularly for applications in cultural tourism. Our approach focuses on building a semantically enriched pipeline capable of transforming unstructured textual artifacts into linked, navigable, and inferentially enhanced digital resources. This enables a new class of digital tourism experiences, where users can interact with history not just as a linear narrative, but as a dynamic network of events, locations, and figures across time. The contributions of this research is a Conceptual Framework to define a unified framework for the integration of NER, domain-specific ontologies, and LLMs in the context of historical text analysis.

II. Proposed Methodology

The ultimate goal is to develop a semantic assistant for cultural tourism, an intelligent system capable of answering complex historical queries, extracting structured data, resolving ambiguities, and presenting information through an engaging interface. This fusion of AI and historical scholarship aims

This research is co-financed by Netcom Engineering S.p.A.

[1]Fabio Clarizia is with the Department of Humanities, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italia, fclarizia@unisa.it

[2]Massimo De Santo is with the Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italia, desanto@unisa.it

[3]Rosario Gaeta is with the Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italia, rgaeta@unisa.it

[4]Rocco Loffredo is with the Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italia, rloffredo@unisa.it

to enhance both academic research and public engagement. The integration of semantic technologies with textual analysis has been explored across various domains, including digital humanities, information retrieval, and intelligent systems but only a limited number of studies have integrated NER [2], [3], ontologies [4], [9], and LLMs into a unified semantic pipeline tailored for cultural tourism. Existing efforts typically focus on one or two of these components. For example, semantic tourism systems often rely on manually curated databases and simple entity linking without deep inference [1], [10]. Conversely, projects in digital humanities may apply LLMs for generative tasks without anchoring the output in ontological models. A key innovation of our work is the combination of these technologies into a closed-loop architecture that supports:

- Entity recognition and disambiguation with domain-adapted models.
- Ontology-based linking and enrichment.
- Contextual reasoning and relation inference using LLMs.

Such an approach not only improves the accuracy of entity linking but also allows the system to generate meaningful answers to complex, tourism-oriented questions. By grounding LLM inference in structured ontologies, we can prevent hallucination and maintain semantic coherence [11], [12]. The proposed approach integrates NER, domain-specific ontologies, and LLMs into a unified pipeline. The goal is to support not only entity extraction and disambiguation but also inferential interpretation and interactive exploration of historical narratives. We divide our methodology into three interconnected components:

- **Construction and adaptation of tourism-specific historical ontologies**: we propose a hybrid ontology that integrates standard schemas (CIDOC CRM, Dublin Core) with domain-specific elements such as historical figures, monuments, rituals, and ancient infrastructure [13]. Built using OWL/RDF and linked to external multilingual resources, the ontology includes temporal-spatial dimensions to resolve name ambiguities [9]. Semantic alignment rules enhance reasoning, allowing LLMs to infer contextual practices, such as associating Venus temples with spring fertility rituals.
- **Domain-adapted NER models for ancient and tourism-related texts**: to improve NER on tourism-related historical texts, we fine-tuned models using a curated corpus from Roman and medieval sources. The pipeline supports both classic and transformer-based models (e.g., RoBERTa, BERT), with manual annotation. Entities were labeled as PER (Historical Persons), LOC (Ancient and Modern Locations),ORG (Guilds, Imperial Institutions), ART (Artifacts, Inscriptions) and EVT (Battles, Festivals, Expeditions).
- **Contextual inference and relation extraction using prompt-engineered LLMs**: while NER extracts entities, LLMs are tasked with understanding their relationships and contextual significance. We leverage GPT-4 and T5 models for zero-shot and few-shot inference using task-specific prompt templates. A successful prompt consists of three parts: Input context (a passage from a historical document), Knowledge seed (facts from the ontology) and Query instruction (task type: relation extraction, question answering, or inference). We mitigate hallucination risks by constraining model outputs to vocabulary drawn from the ontology and validating relations through a consistency-check module [11].

TABLE I
TECHNICAL STACK AND IMPLEMENTATION

Component	Technology Used
OCR & Text Normalization	Tesseract, GROBID, spaCy pipeline
NER	RoBERTa fine-tuned model, spaCy custom
Entity Linking	DBpedia Spotlight, Wikidata API
Ontology Store	GraphDB (RDF4J), Protégé, OWL 2.0
Inference Engine	GPT-4 API, T5, prompt engine (Python)
UI Layer	React.js, D3.js for graph, Leaflet for maps

III. CONCLUSION

The modular design ensures each component can be updated or replaced without breaking downstream processes. All outputs are stored in a triplestore, allowing querying through SPARQL or natural language templates. The preliminary results demonstrate that the integration of NER, ontologies, and LLMs presents a powerful methodology for extracting, linking, and reasoning over historical information in tourism-relevant texts.

REFERENCES

[1] C. F. R. et al., "Semantic technologies for cultural tourism: A review," *Information*, vol. 13, no. 1, p. 9, 2022.

[2] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[3] A. Choubey, R. Pandey, and A. Shrivastava, "Named entity recognition: A literature review," *International Journal of Computer Applications*, vol. 181, no. 32, pp. 20–25, 2019.

[4] A. Gangemi, "Ontology design patterns for semantic web content," in *ISWC*, ser. LNCS, vol. 3729, 2005, pp. 262–276.

[5] P. Buitelaar and T. Eigner, "Ontology-based semantic annotation of historical texts," in *LREC*, 2006.

[6] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.

[7] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. [Online]. Available: https://aclanthology.org/Q19-1016/

[8] A. V. et al., "Attention is all you need," in *NeurIPS*, 2017.

[9] S. Borgo and C. Masolo, "Ontological foundations of dolce," in *Theory and applications of ontology: Computer applications*. Springer, 2010, pp. 279–295.

[10] Á. García-Crespo, J. L. López-Cuadrado, R. Colomo-Palacios, I. González-Carrasco, and B. Ruiz-Mezcua, "Sem-fit: A semantic based expert system to provide recommendations in the tourism domain," *Expert systems with applications*, vol. 38, no. 10, pp. 13 310–13 319, 2011.

[11] A. Gaur, J. Li, and H. Ji, "Knowledge-grounded generation with structured semantic memory," in *ACL*, 2021.

[12] J. Chen, X. Yan, and R. Jia, "Triple generation for open-domain knowledge bases with llms," in *arXiv:2311.01562*, 2023.

[13] M. Zarka and J. D. Fernandez, "Semantic tourism data integration using linked data," *SEMANTiCS*, 2018.