

From Unstructured Heritage Texts to Validated Knowledge Graphs: A NLP-Ontology-LLM Pipeline with Ground-Truth-Free Evaluation

Fabio Clarizia^{1,*}, Massimo De Santo², Pasquale Esposito¹, Rocco Loffredo² and Benedetta Sabini¹

¹Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, Italy

²Hermes Lab, Department of Humanities, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, Italy

Abstract

This paper presents a hybrid framework for the semantic enrichment of cultural heritage texts, integrating Named Entity Recognition (NER), domain ontologies, and Large Language Models (LLMs) to transform unstructured historical narratives into structured, interoperable knowledge graphs. The pipeline combines CIDOC-CRM-aligned entity extraction, a two-layer nominal/verbal ontology alignment, and prompt-engineered LLM inference, producing RDF-compliant graphs validated through SHACL constraints. To assess the quality of the automatically generated knowledge graph without a manually curated reference ontology, we introduce a ground-truth-free evaluation framework comprising five complementary modules: structural metrics, annotation coverage, schema richness following the Onto-QA methodology, graph connectivity via the Average Degree Score, and an LLM-as-Judge protocol. Applied to a knowledge graph of 96,348 triples encoding 12,563 individuals derived from 944 Italian library records, the evaluation confirms that the baseline pipeline produces domain-coherent and structurally sound graphs, achieving an overall quality score of 8.0/10, with strong naming conventions (9.0/10) and domain coherence (9.0/10). Future work will focus on automated annotation generation, OWL 2 DL consistency enforcement, and scaling the pipeline to multilingual and cross-domain corpora.

Keywords

Named Entity Recognition, Large Language Models, Cultural Heritage, Knowledge Graphs, Ontology Engineering, Digital Humanities, CIDOC CRM, Semantic Web

1. Introduction

In an era of major technological transformation, advances in digital tools offer a significant opportunity to completely rethink our approach to historical texts. This need stems from the desire to fulfill the cultural and scientific objective of improving the accessibility and systematic searchability of complex corpora – which are often unannotated and housed in rapidly expanding, unstructured digital archives. The progressive digitization of cultural heritage has led to the proliferation of digital archives that are often heterogeneous and not designed according to shared standards. These collections, therefore, remain difficult to query and reuse effectively, particularly for those without specific scientific expertise. Transforming historical narratives into machine-interpretable knowledge remains a central challenge due to their semantic ambiguity, diachronic variation and reliance on implicit references. This problem is particularly evident when relevant entities and relationships are absent from existing knowledge bases. Although Named Entity Recognition (NER) provides an initial level of structuring, it is insufficient to capture the deeper semantic relationships inherent in such texts. To address this issue, the proposed approach combines formal ontologies – in particular CIDOC CRM – to provide an explicit semantic foundation, with large language models (LLMs) to infer implicit relationships; their outputs are validated through a *ground-truth-free* evaluation strategy. Within this framework, we develop an integrated pipeline incorporating embedding-based similarity, rule-based alignment,

WOA 2026, the 27th Workshop From Objects to Agents Salerno, June 15-17, 2026

✉ fclarizia@unisa.it (F. Clarizia); desanto@unisa.it (M. D. Santo); pasesposito@unisa.it (P. Esposito); rloffredo@unisa.it (R. Loffredo); bsabini@unisa.it (B. Sabini)

🆔 0000-0003-2660-6759 (F. Clarizia); 0000-0002-8671-4869 (M. D. Santo); 0009-0006-3464-5861 (P. Esposito); 0009-0007-0799-1168 (R. Loffredo); 0009-0004-0273-2931 (B. Sabini)



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

and ontology-guided inference. This combination enables entity disambiguation, relation extraction, and semantic validation, fostering the transformation of unstructured historical texts into coherent, FAIR-compliant, and context-aware knowledge representations.

2. Problem Statement

The fundamental prerequisite for the semantic enrichment of CH is the transformation of the unstructured historical sources into structured data. NER serves as a baseline for the detection of the entities, such as people, places, artifacts and dates. Nevertheless, is insufficient by itself for the resolution of semantic overlaps and fails to extract the latent connections necessary for deep cultural analysis. This work addresses several challenges. When dealing with historical and cultural texts, a single lexeme may refer to different entities due to semantic variations across diachrony and synchrony; therefore, the first obstacle is the need to disambiguate the terms within the text. Another challenge involves the necessity of mapping entities to specific classes and instances that comply with current standards to ensure interoperability. Further issues concern latent connections—namely, the historical associations implicit in the narrative—which require advanced inference to be formalized into structured data. Finally, there is the need to mitigate the tendency of LLMs to ‘hallucinate’ and produce inaccuracies.

The aim of this work is to formulate an integrated pipeline that surmounts these challenges by combining entity extraction, ontology-driven disambiguation, and LLM inference.

3. State of the Art

3.1. Semantic Enrichment in Cultural Heritage

One of the main researched areas within contemporary Natural Language Processing, is the application of NER systems for their crucial role in distilling structured data from unstructured texts.[1]. However, deploying NER on historical archives introduces distinct complexities in terms of orthographic inconsistencies, archaic spelling conventions and outdated linguistic patterns [2]. To tackle these issues, recent studies have increasingly focused on tailoring transformer-based models, like BERT and DistilBERT, to specialized corpora [3].

To bridge the gap between raw entity extraction and structured knowledge representation, the information retrieved through NER must be organized within formal frameworks. To achieve this semantic interoperability and facilitate complex querying across heterogeneous cultural heritage datasets, the implementation of ontologies has become a standard practice. The CIDOC Conceptual Reference Model (CIDOC CRM) acts as the prevailing framework for organizing archival, library and museum data, with its robustness and widespread acceptance validated by systematic literature [4]. This model is frequently paired with lightweight metadata standards, such as Dublin Core, for web applications, while domain-specific extensions are utilized to capture nuanced information like spatial narratives and intangible cultural assets. The critical function of linked data within the cultural heritage sector is further evidenced by major semantic infrastructures. Collaborative platforms like SampoSampo [5] successfully combine linked open data with detailed ontological models, empowering users to explore and uncover new knowledge across decentralized repositories.

3.2. Knowledge-Grounded Inference

LLMs have fostered the development of hybrid methodologies that merge statistical inference with structured knowledge, obtaining impressive results in relation extraction, contextual reasoning and summarization. Within the cultural heritage domain, contemporary research is investigating the fusion of LLMs and formal ontologies to facilitate the creation of knowledge graphs and ensure semantic validation [6]. For instance, Wang and Zhang introduced a CIDOC CRM-based workflow that utilizes LLMs to extract entities and their relations while strictly adhering to ontological boundaries [7]. On a larger scale, digital humanities literature highlights the use of ontology-driven prompting and

knowledge-grounded inference to balance generative versatility with formal semantic accuracy [8]. Guaranteeing that extracted relations accurately reflect explicit domain semantics is still a significant ongoing challenge, especially when navigating the ambiguities inherent in historical narratives [9]. LLMs are prone to "hallucinations" when they must make inferences about complex structures such as ontologies and experience a decline in performance when they lack guided examples or explicit structural constraints [10]. Given the absence of a manually curated reference ontology for the target cultural heritage corpus, ensuring the reliability of the integrated system requires a non-comparative validation paradigm. To this end, we implement a robust, ground-truth-free evaluation strategy structured around five complementary modules. This architecture systematically assesses structural quality, metadata metadata coverage, schema richness, and semantic coherence, providing a comprehensive quality control loop that guarantees the formal integrity of the generated knowledge graph without reference-set dependencies.

4. Methodology

In order to convert the unstructured cultural heritage narratives into a structured and machine-readable format, both ensuring the accuracy and the narrative depth of the analyzed texts, the proposed methodology follows a multi-phase pipeline that integrates NLP techniques, symbolic ontology design and the reasoning power of LLMs. The complete architectural framework is depicted in Figure 1 (Section ??).

4.1. System Architecture Overview

The proposed system integrates symbolic and neural components into a layered semantic enrichment pipeline, as outlined below:

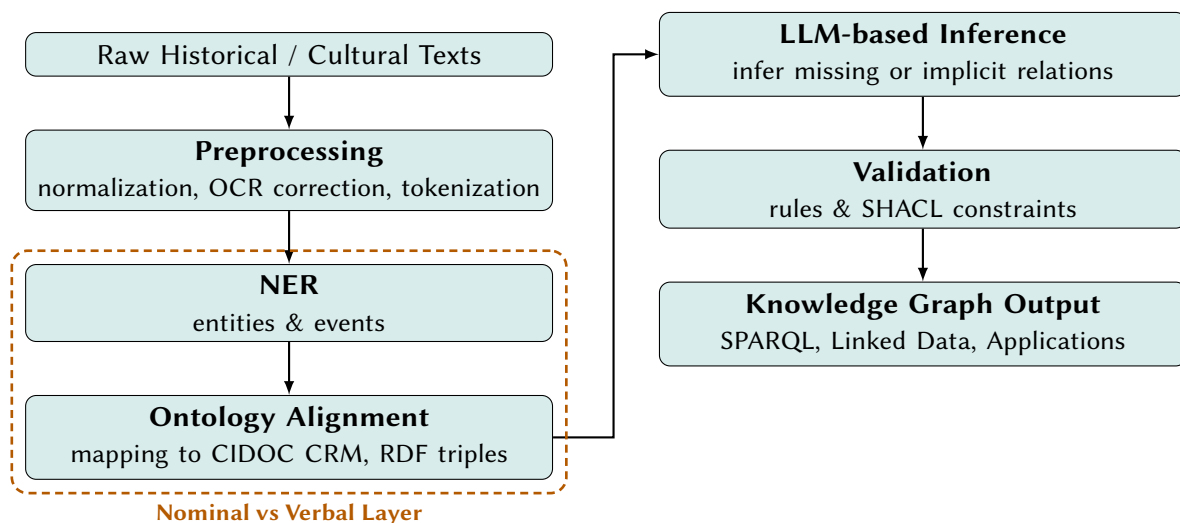


Figure 1: Conceptual representation of the proposed framework: unprocessed text documents are analyzed through a series of consecutive steps. This workflow transforms the raw material into structured knowledge graphs tailored to the cultural heritage domain.

The proposed framework implements a five-step semantic enrichment process, illustrated in Figure 1, that transforms raw cultural heritage texts into structured and validated knowledge graphs.

The first phase involves preparing raw cultural and historical heritage texts through cleaning, normalization, and tokenization. Once the OCR transcription of the texts has been obtained, we proceed to correct digitization errors, optimizing the input data and ensuring the accuracy of subsequent natural language processing stages, which will receive standardized, “noise-free” input. To examine this process in greater detail, the following subsections describe the specific architectures and parameters employed.

4.2. Knowledge Base Construction

Named Entity Recognition. The preprocessed text is analyzed by fine-tuned NER architectures to extract relevant entities and events. This stage identifies the nominal elements within the historical record (such as people, places, dates and so on). The goal here is to answer basic investigative questions without yet defining the complex semantic relationships between the detected elements. This stage involves the detection and categorization of entities within tourism and historical corpora, where NER serves as the critical mechanism for generating the ontology’s baseline data. This extraction is governed by the structural separation of the **nominal layer** (entities and nodes) from the **verbal layer** (relations). Thus, the NER step specifically targets nominal elements—including people, places, historical artifacts, dates, and events—destined to become ontological nodes. To maintain semantic interoperability with cultural heritage protocols, a supervised filtering approach driven by the CIDOC CRM ontology was adopted. During this preliminary stage, NER categories were mapped onto a refined taxonomy of 99 distinct CIDOC CRM classes (e.g., E21_Person, E53_Place, E22_Human-Made Object), alongside optional URIs for Linked Data integration. Although this high level of granularity ensured precise classification, it inadvertently resulted in structural fragmentation and redundancy across overlapping semantic categories.

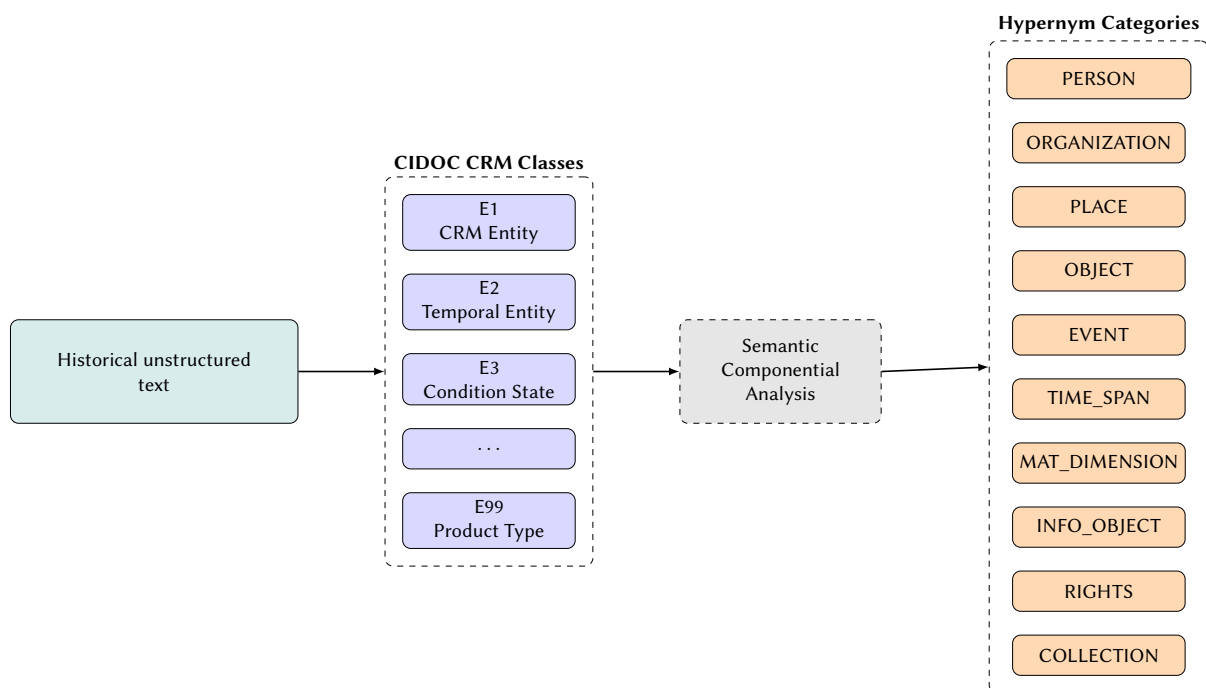


Figure 2: A two-stage mapping framework for NER: extracted labels are initially filtered and aligned using fine-grained CIDOC CRM classes, and subsequently consolidated into broader categories via semantic componential analysis to facilitate practical extraction tasks and ontology integration.

To optimize both practical annotation workflows and model training, the preliminary mapping of 99 CIDOC CRM classes was aggregated into a more concise set of hypernymic categories. This consolidation was driven by semantic componential analysis [11, 12], which served as a rigorous framework for grouping CIDOC classes that exhibit common foundational semantic traits. By doing so, the approach effectively mitigates label fragmentation while maintaining the crucial distinctions required for subsequent ontology integration. Ultimately, this yields a manageable label space for the NER module, ensuring that every predicted category retains a clear and direct link back to its foundational CIDOC-aligned structure.

The definitive set consists of:

- PERSON – individuals involved (artists, collectors, restorers).

- ORGANIZATION – institutions such as museums, foundations, auction houses, laboratories.
- PLACE – geographic locations, buildings, and physical features (e.g., “Alps”, “Sala degli Arazzi”).
- OBJECT – human-made artifacts (paintings, statues, coins, furniture).
- EVENT – relevant events (creation, exhibitions, restorations, excavations, donations).
- TIME_SPAN – dates or historical periods (e.g., “15th century”, “April 12, 2024 June 30, 2024”).
- MATERIAL_DIMENSION – materials, techniques, measures, weights (e.g., “oil on canvas”, “h 75 cm”).
- INFORMATION_OBJECT – related information resources (catalog entries, photographs, 3D models, PDFs).
- RIGHTS – legal and economic aspects (ownership, licenses, contracts of sale).
- COLLECTION – curated holdings or reference sets (“Medici Collection”, “Rossi Archive Fund”).

As illustrated in Figure 2, this dual-phase filtering approach enabled us to retain the semantic depth of the CIDOC CRM while keeping the NER task computationally manageable and cross-compatible. By anchoring the category consolidation in both linguistic principles (componential semantics) and practical computational needs (clustering for model optimization), the extracted entities map to specific ontology nodes. These entities thus establish the nominal groundwork necessary for the subsequent formulation of relationships (the verbal layer).

Ontology Alignment. This stage converts surface-level detections into a formal representation by aligning entities with a multi-layered ontology. The model blends CIDOC CRM, Dublin Core and domain-specific vocabularies to cover the full spectrum of heritage, from monuments to historical events. The core of this process lies in the separation of the *nominal* and *verbal* layers. Mapping is executed via a mix of string normalization and embedding-based similarity scores. In cases of high ambiguity the framework triggers a disambiguation protocol. This protocol integrates rule-based logic and expert dictionaries with geographic and temporal constraints to ensure precise entity resolution. Through the ontology, we were able to identify both the tangible and intangible aspects of heritage (collections, historical events, monuments, and rituals), enabling rich semantic integration across diverse sources. The use of the ontological infrastructure made it possible to map both the tangible and intangible components of heritage (such as collections, historical events, monumental complexes, and ritual practices), thereby fostering deep semantic integration among heterogeneous information sources.

The structure of our research is based on a clear distinction between the **nominal layer** and the **verbal layer**:

- The **nominal layer** identifies conceptual nodes classifying, for example, *Alexandria* as `E53_Place`.
- The **verbal layer** establishes the interconnections between nodes (e.g., *Excavation carriedOutBy Fiorelli*).

This dichotomy is essential: isolated entities serve merely as referential anchors, but it is the relational interweaving that gives them historical and semantic depth.

The Entity–ontology alignment proceeds along two lines:

1. **Surface lexical correspondence:** uses string-matching algorithms and text normalization procedures to establish a direct link between mentions and the canonical labels of the ontology.
2. **Semantic embedding:** leverages the calculation of contextualized embeddings for both extracted entities and candidate ontology classes, refining the alignment by evaluating deep semantic proximity.

In instances where referential ambiguities arise (such as toponymic homonyms, e.g. "Alexandria"), the disambiguation process operates synergistically by combining rule-based inference, the use of expert-validated domain vocabularies, and systematic cross-referencing with spatial and temporal metadata.

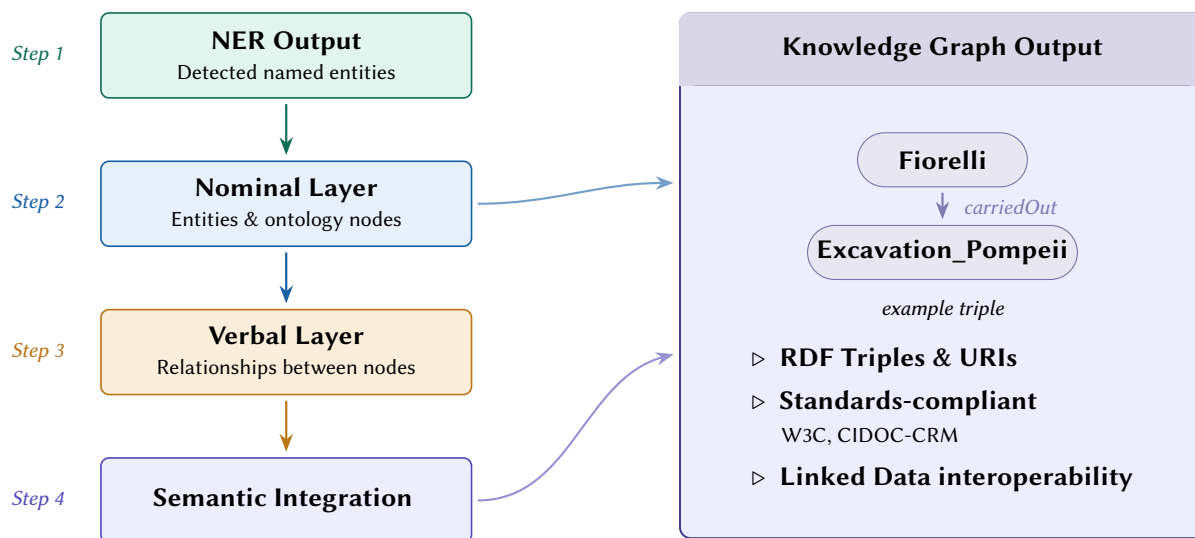


Figure 3: Ontology generation workflow. Entities extracted via NER constitute the nominal layer (graph nodes), which are subsequently linked by the verbal layer (relational edges). This combined structure is then serialized into URI-enriched RDF triples, yielding a fully machine-interpretable knowledge graph.

Knowledge Graph generation. As illustrated in Figure 3, the pipeline’s culminating phase involves converting the conceptual ontology into a machine-actionable knowledge graph. To bridge the gap between the ontology and the final KG, the system generates CIDOC-CRM-compliant RDF triples, securing broad interoperability through targeted implementation strategies:

1. Entities receive persistent identifiers (URIs), which are cross-referenced with established linked-data resources whenever feasible to foster data reuse and resolve referential ambiguities.
2. To ensure accurate downstream analysis, the graph maintains a strict separation between object properties (entity-to-entity relations) and data properties (attributes expressed as typed literals, such as dates, materials, or dimensions).
3. Every generated export explicitly declares itself as an ontology and integrates the CIDOC CRM schema, allowing external platforms to seamlessly comprehend its conceptual structure and boundaries.
4. To support open publication and seamless linked open data ecosystem integration, the resulting KG can be output in widely accepted serialization formats, including JSON-LD, Turtle and RDF/XML.

4.3. Large Language Model Integration

After aligning entities and primary relations with the ontology, LLMs (such as GPT or LLaMA) are utilized to deduce implicit connections not overtly mentioned in the source text. Serving as a layer for semantic enrichment, this LLM-driven process expands the knowledge graph by identifying nuanced historical and narrative links that elude basic extraction methods. This phase focuses on the verbal layer, augmenting the ontology’s symbolic framework while adhering strictly to its defined semantic constraints. Interaction with the LLM occurs through prompt engineering techniques, structuring each request around three fundamental pillars:

1. **Input Context:** the original text fragment to be analyzed.
2. **Knowledge Seed:** a selection of related entities or relevant ontological data.
3. **Query Instruction:** the operational command, such as the request to determine the historical period or political jurisdiction of a given event.

This query scheme limits the generation of incorrect information (hallucinations) by forcing the model to rely exclusively on the provided data. The outputs are subsequently filtered and refined to identify potential logical links (e.g., *Person visited Place on Date*). Finally, once a certain reliability threshold is exceeded, these relationships are integrated into the Knowledge Graph. The core integration tasks within this pipeline were driven by LLM such as Qwen 3.6 and GPT-OSS.

Validation. The final stage implements rigorous verification protocols to guarantee the semantic integrity of the augmented knowledge graph. Using SHACL constraints and rule-based logic, the system enforces ontological alignment, identifies contradictions, and refines inferred relationships. This iterative validation loop allows for continuous improvement by feeding corrections back into the inference layer before finalization. The resulting high-fidelity knowledge graph is then available for SPARQL querying, Linked Open Data publication, or integration into cultural heritage and tourism platforms. The system’s architecture is illustrated in Figure 1. The right side delineates the core symbolic workflow, tracing the transformation of raw text into ontology-aligned structures. Conversely, the left side highlights the semantic enrichment layer, where LLM-based inference and validation act as auxiliary components to expand and iteratively polish the structured knowledge. This process is formalized in the simplified pseudocode of Algorithm 1.

Algorithm 1 Semantic Enrichment Workflow

- 1: **Input:** Unstructured text corpus T
 - 2: Preprocess T (normalization, OCR correction, tokenization)
 - 3: Apply neural models to extract entities E and events V
 - 4: Map E and V to ontology classes and instances
 - 5: Infer relations R using prompt-engineered LLMs
 - 6: Apply rule-based reasoning to refine (E, V, R)
 - 7: Instantiate RDF triples (s, p, o) from (E, V, R)
 - 8: Validate triples using SHACL constraints
 - 9: **Output:** Knowledge graph KG
-

The framework outputs an enriched knowledge graph that can be queried using SPARQL and integrated into semantic search systems or tourism and cultural applications. Thanks to its modular design, it is possible to update NER models, ontologies, or LLM configurations individually. The architecture is based on the distinction between the nominal level (entity nodes) and the relational level (events and inferred connections). This hybrid approach combines the flexibility of neural models with the rigor of symbolic reasoning, ensuring data consistency and traceability. The resulting knowledge graphs are designed for advanced querying, provenance tracking, and large-scale integration into digital scholarship ecosystems. In summary, the methodology integrates three pillars:

- NER: for initial structured annotations.
- Ontologies: for semantic consistency and disambiguation.
- LLMs: for inferential capabilities.

The result is a robust system that transforms unstructured text into context-rich, interoperable graphs.

5. Evaluation

To assess the quality of the knowledge graph produced by the proposed pipeline, we adopt a *ground-truth-free* evaluation strategy. Since no manually curated reference ontology is available for the target corpus, we rely on a multi-module framework that measures structural quality, annotation coverage, schema richness, graph connectivity, and holistic semantic coherence without requiring a gold-standard counterpart [13].

5.1. Evaluation Framework

The evaluation pipeline integrates five complementary assessment modules, each targeting a distinct quality dimension.

Structural metrics. We count the core ontological elements of the generated KG (classes, individuals, object and datatype properties, and their taxonomic relations) and verify whether each element is properly declared according to OWL 2 conventions [14].

Annotation coverage. For each class and property we check the presence of human-readable metadata, namely `rdfs:label` and `rdfs:comment` annotations for classes, and `rdfs:domain/rdfs:range` declarations for object properties. Low annotation coverage signals reduced interpretability and hinders interoperability with external systems that rely on natural-language metadata [15].

Schema richness. We compute four richness metrics from the Onto-QA framework [16]:

- **Relationship Richness (RR):** $|OP| / (|OP| + |SC|)$ – proportion of non-taxonomic over all relations.
- **Attribute Richness (AR):** $|DP| / |C|$ – average number of datatype properties per class.
- **Inheritance Richness (IR):** $|SC| / |C|$ – depth and branching of the class hierarchy.
- **Class Richness (CR):** $|C_{\text{inst}}| / |C|$ – fraction of classes with at least one individual.

where $|OP|$, $|DP|$, $|SC|$, $|C|$, and $|C_{\text{inst}}|$ denote object properties, datatype properties, `subClassOf` axioms, classes, and instantiated classes, respectively.

Graph connectivity. We compute the Average Degree Score D_{avg} introduced by Guéret et al. [17] and adopted by Bakker et al. [13]:

$$D_{\text{avg}} = \frac{|I| + |C|}{|P_{II}| + |P_{IC}| + |P_{IL}|} \quad (1)$$

where $|I|$ is the number of individuals, $|P_{II}|$ the individual-to-individual assertions, $|P_{IC}|$ the individual-to-class assertions, and $|P_{IL}|$ the individual-to-literal assertions. The metric expresses the average number of graph elements reachable per assertion; a low score signals sparse interconnection between individuals.

LLM-as-Judge. Following recent work on automated ontology evaluation [18], we prompt LaMA 3.1 8B via Ollama with a structured verbalization of the ontology schema and ask it to rate five dimensions on a 0–10 scale: *completeness*, *naming quality*, *structural quality*, *domain coherence*, and *overall quality*. The model is instructed to return a structured JSON object with per-dimension scores and justifications, using a temperature of 0.1 to minimize stochasticity. The deployment of LLaMA 3.1 8B model is justified by experimental reproducibility and environmental sustainability, ensuring identical judging conditions over time and minimizing the computing costs.

5.2. Experimental Setup

The evaluation was conducted on `bu11000.rdf`, a knowledge graph automatically generated by the proposed pipeline from a corpus of 944 Italian-language library records with bilingual (IT/EN) abstracts. Each record was processed through the full pipeline (preprocessing, NER, ontology alignment, SRL-based verbal layer extraction, and RDF serialization) yielding a graph that encodes both bibliographic metadata and the semantic role annotations extracted from the abstracts. The graph was serialized in RDF/XML format and parsed using `rdflib` [19]. The LLM judge was executed locally via Ollama without external API dependencies.

5.3. Results

Structural profile. Table 1 summarises the ontological profile of the evaluated KG. The schema encodes 15 OWL classes distributed across two namespaces: three application-specific classes (`ex:Book`, `ex:Abstract`, `ex:RelationAssertion`) and twelve classes derived from the SRL ontology (`srl:ENTITY` hierarchy). The 96,348 triples encode 12,563 individuals, of which 8,445 are entity mentions, 1,977 are semantic role assertions, 944 are book records, and 1,197 are bilingual abstracts.

Table 1

Ontological profile of the evaluated knowledge graph (`bu11000.rdf`).

Category	Element	Count
Schema (TBox)	OWL classes	15
	Object properties	18
	Datatype properties	18
	<code>rdfs:subClassOf</code> axioms	10
	Max. hierarchy depth	1
Data (ABox)	Individuals	12,563
	Total triples	96,348

Annotation coverage. Table 2 reports annotation coverage across the schema. Object property signatures are well-specified (`rdfs:domain` at 100%, `rdfs:range` at 89%).

Table 2

Annotation coverage of the generated ontology. Higher values indicate greater human-readability.

Annotation	Coverage
Object properties with <code>rdfs:domain</code>	100%
Object properties with <code>rdfs:range</code>	89%

Schema richness. Table 3 presents the Onto-QA richness metrics [16]. The Attribute Richness ($AR = 1.20$) confirms that each class is associated on average with more than one datatype property, indicating rich data modeling. The Class Richness ($CR = 0.933$) shows that nearly all classes are populated with individuals, reflecting good domain coverage. The Relationship Richness ($RR = 0.643$) and Inheritance Richness ($IR = 0.667$) indicate a balanced distribution between non-taxonomic relations and hierarchical structure, in line with ontologies designed for knowledge graph applications rather than purely taxonomic classification.

Graph connectivity. Table 4 reports the connectivity metrics. The ABox degree score $D_{avg} = 0.253$ indicates that each assertion connects on average approximately four ontological elements, computed over $P_{II} = 24,248$ individual-to-individual assertions and $P_{IL} = 30,527$ individual-to-literal assertions

Table 3

Schema richness metrics (Onto-QA framework [16]). RR: Relationship Richness; AR: Attribute Richness; IR: Inheritance Richness; CR: Class Richness.

Metric	Value	Formula
Relationship Richness (RR)	0.643	$ OP / (OP + SC)$
Attribute Richness (AR)	1.200	$ DP / C $
Inheritance Richness (IR)	0.667	$ SC / C $
Class Richness (CR)	0.933	$ C_{inst} / C $

(with $P_{IC} = 0$, as no property directly links individuals to classes). The TBox density of 0.333 reveals that only one third of the classes participate actively in object property signatures, while 26.7% of classes appear as isolated nodes with no domain or range involvement.

Table 4

Connectivity metrics of the generated knowledge graph. D_{avg} follows [17, 13].

Metric	Value	Detail
D_{avg} (ABox degree)	0.253	$P_{II}=24,248$; $P_{IC}=0$; $P_{IL}=30,527$
TBox density	0.333	classes in OP signatures
Classes as <code>rdfs:domain</code>	20.0%	3 of 15
Classes as <code>rdfs:range</code>	26.7%	4 of 15
Isolated classes	26.7%	4 of 15

LLM-as-Judge. Table 5 reports the automated evaluation scores produced by LLaMA 3.1 8B [20, 18]. The model assigned high scores to naming quality (9.0/10) and domain coherence (9.0/10), confirming that the pipeline respects standard OWL naming conventions and produces semantically coherent content. The completeness score (8.0/10) reflects a partial absence of relational context for some individuals, while the structural quality score (7.0/10) flags the shallow hierarchy and the presence of isolated classes as areas for improvement. The overall score of 8.0/10 indicates that the baseline system produces structurally sound and domain-coherent knowledge graphs, while also identifying concrete targets for the next development cycle.

Table 5

LLM-as-Judge evaluation scores (0–10 scale) produced by LLaMA 3.1 8B [20] on a structured verbalization of the ontology schema ($T = 0.1$). Higher scores are better.

Dimension	Score	Key observation
Completeness	8.0	Concepts captured; some individuals lack relational context.
Naming quality	9.0	CamelCase/camelCase conventions respected throughout.
Structural quality	7.0	Well-formed hierarchy; isolated classes reduce connectivity.
Domain coherence	9.0	Semantically coherent and domain-appropriate content.
Overall	8.0	Solid baseline; targeted refinements required.

6. Discussion of Results

The evaluation reveals that the pipeline successfully produces large-scale, domain-coherent knowledge graphs with well-defined property signatures and rich data modeling (CR = 0.933, AR = 1.20).

The area for improvement concerns the 26.7% of isolated classes. These correspond primarily to the more abstract top-level nodes of the SRL hierarchy (`srl:ENTITY`, `srl:COLLECTION`, `srl:MATERIAL_DIMENSION`, `srl:RIGHTS`) that function as conceptual anchors but are not currently referenced in any object property's domain or range. Introducing bridging axioms between the `ex:` and `srl:` namespaces, or extending the property signatures of these classes, would improve both TBox density and overall connectivity.

These findings confirm that the proposed pipeline establishes a functional and semantically coherent baseline while providing actionable targets for the next development cycle, consistent with the improvement directions outlined in Section 7.

6.1. Computational Impact and Scalability Analysis

Although designed with a modular approach, the pipeline's actual end-to-end scalability is constrained by the number of entities identified during the NER phase and the frequency of LLM calls used to extract relationships (Algorithm 1, steps 3-8). Inference using LLM incurs costs related to the tokens processed and increases overall latency; these parameters scale proportionally to the size of the context window, the volume of the "knowledge seed", and the number of prompts per text. issued per document. A practical measure to mitigate this impact consists of limiting calls to the model: by first completing the NER and ontology alignment processes, the LLM is queried only as a last resort, specifically in cases of referential ambiguity and when the rules do not guarantee high reliability in identifying relationships. In this way, the rate of prompt generation is reduced from "one per mention" to "one per anomaly," ensuring sub-linear scalability relative to the total volume of data processed.

6.2. Expected Bottlenecks and Practical Mitigation

During implementation, the architecture presents four potential bottlenecks that require specific measures:

1. **OCR Preprocessing and Normalization:** The degraded quality of historical documents can generate errors that affect the Named Entity Recognition (NER) phase, producing false positives. Adopting a lightweight syntactic correction module, integrated with domain-specific dictionaries, is the most efficient solution for stabilizing the output.
2. **Entity Resolution and Ontological Alignment:** Referential ambiguity increases proportionally to the size of the candidate space (URIs/classes). It is necessary to apply pruning filters (based on string, temporal, or geographic criteria) before calculating semantic similarity via embeddings, while also utilizing an alias table to manage recurring entities.
3. **Inference via LLM:** Processing latency and API restrictions can limit system throughput. To optimize performance, a gating mechanism must be implemented (limiting model use to unresolved steps only), the context must be segmented into smaller chunks, and structured output formats must be enforced to minimize the need for manual review.
4. **Knowledge Graph Validation:** Checking constraints on large graphs entails high computational costs. It is preferable to perform incremental validations exclusively on newly inserted triples and to discard malformed relationships in advance, thereby avoiding costly ex-post structural repairs.

These measures ensure the framework's operability and enable independent scalability of each component module.

7. Conclusion and Future Work

The framework presented in this work proposes a hybrid approach integrating NER, domain-specific ontologies, and LLMs to enhance the semantic enrichment of cultural heritage documents and construct a robust and reliable knowledge infrastructure for agents. The proposed framework demonstrates the conversion of unstructured historical sources, including manuscripts, epigraphy, and travelogues, into structured, machine-readable repositories, effectively bridging the gap between advanced scholarly inquiry, digital tools for cultural tourism and AI-driven Information Retrieval. Crucial to our approach is the operational distinction between the nominal layer (entity creation) and the verbal layer (relationship definition), ensuring a modular construction of the ontological domain. This structural decoupling ensures that entities are robustly grounded and disambiguated prior to the insertion of semantic predicates, thereby enhancing both the logical consistency and the interpretability of the generated knowledge graphs. This infrastructure serves as an ideal substrate for cognitive agents. Building upon this architecture, future research will focus on:

- Developing automated techniques for the identification and classification of complex relational predicates between entities to enrich the verbal layer.
- Implementing robust validation protocols for ontology population, benchmarking results against a curated ground truth and integrating hybrid approaches that combine rule-based constraints with expert-in-the-loop verification.
- Scaling the pipeline to multilingual and cross-domain corpora to evaluate its generalizability and impact across broader digital humanities and cultural heritage frameworks.

Through these developments, we aim to advance ontology-driven retrieval and reasoning, further integrating the framework into RAG architectures to enhance both knowledge accessibility and the explainability of generated insights.

Acknowledgments

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

8. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Text Translation, Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Linguisticae Investigationes* 30 (2007) 3–26.
- [2] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Diachronic evaluation of ner systems on old newspapers, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [4] P. Ferreira-Lopes, E. González-Gracia, Performing a systematic literature review on the implementation of the cidoc crm in cultural heritage, *Journal on Computing and Cultural Heritage* 18 (2025). doi:10.1145/3771098.

- [5] E. Hyvönen, P. Leskinen, A. Ahola, H. Rantala, J. Tuominen, Samposampo: A portal for studying enriched data and semantic connections on a cultural heritage linked open data cloud, in: *Lecture Notes in Computer Science*, volume 15832, 2026, pp. 67–74. doi:10.1007/978-3-031-99554-5_13.
- [6] F. Clarizia, M. De Santo, R. Gaeta, R. Loffredo, Enhancement large language models domain through ontology-based retrieval-augmented generation, *International Journal on Semantic Web and Information Systems (IJSWIS)* 21 (2025) 1–29. doi:10.4018/IJSWIS.392507.
- [7] Y. Wang, M. Zhang, Cidoc crm-based knowledge graph construction for cultural heritage using large language models, *Applied Sciences (Switzerland)* 15 (2025). doi:10.3390/app152212063.
- [8] E. Hyvönen, Digital humanities on the semantic web: from infrastructure to practical applications, ai-based knowledge discovery, and web of wisdom, in: *Lecture Notes in Computer Science*, volume 16097, 2026, pp. 3–8. doi:10.1007/978-3-032-05409-8_1.
- [9] F. Colace, G. D’Aniello, M. De Santo, R. Gaeta, G. Zuchtriegel, Situation-aware cyber-physical-social system for cultural heritage, *Intelligent Systems with Applications* 27 (2025) 200544. URL: <https://www.sciencedirect.com/science/article/pii/S2667305325000705>. doi:<https://doi.org/10.1016/j.iswa.2025.200544>.
- [10] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, 2024. URL: <https://arxiv.org/abs/2305.13168>. arXiv:2305.13168.
- [11] J. J. Katz, J. A. Fodor, The structure of a semantic theory, *Language* 39 (1963) 170–210. doi:10.2307/411200.
- [12] J. Lyons, *Semantics*, volume 2, Cambridge University Press, Cambridge, 1977.
- [13] R. M. Bakker, D. L. Di Scala, M. H. T. de Boer, Ontology learning from text: An analysis on LLM performance, in: *Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation (NLP4KGC)*, co-located with SEMANTiCS 2024, Amsterdam, The Netherlands, 2024.
- [14] F. Baader, I. Horrocks, C. Lutz, U. Sattler, *An Introduction to Description Logic*, Cambridge University Press, 2017. doi:10.1017/9781139025355.
- [15] R. S. I. Wilson, J. S. Goonetillake, A. Ginige, W. A. Indika, Ontology quality evaluation methodology, in: *International Conference on Computational Science and Its Applications (ICCSA 2022)*, *Lecture Notes in Computer Science*, Springer, 2022, pp. 509–528. doi:10.1007/978-3-031-10522-7_35.
- [16] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, B. Aleman-Meza, OntoQA: Metric-based ontology quality analysis, in: *Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources (KADAHTTP)*, co-located with ICDM, 2005.
- [17] C. Guéret, P. Groth, C. Stadler, J. Lehmann, Assessing linked data mappings using network measures, in: *The Semantic Web: Research and Applications. Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012)*, volume 7295 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 87–102. doi:10.1007/978-3-642-30284-8_11.
- [18] S. Tsaneva, S. Vasic, M. Sabou, LLM-driven ontology evaluation: Verifying ontology restrictions with ChatGPT, in: *The Semantic Web: ESWC 2024 Satellite Events*, *Lecture Notes in Computer Science*, Springer, 2024.
- [19] RDFLib Team, RDFLib: A Python library for working with RDF, 2024. URL: <https://rdflib.readthedocs.io>, version 7.0, <https://github.com/RDFLib/rdflib>.
- [20] A. Grattafiori, A. Dubey, A. Jauhri, et al., The LLaMA 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).