From Text to Triples: Large Language Models for Ontology-Aligned Annotation in Archaeology

Ali Hariri[1]

[1]ISAE-ENSMA, LIAS Laboratory, France

Abstract

Semantic annotation of archaeological documentation in line with the CIDOC Conceptual Reference Model (CIDOC CRM) remains labour-intensive and demands specialised expertise. This doctoral research explores how Large Language Models (LLM) can alleviate that burden through a two-phase agenda: first, steering general-purpose models with curated CIDOC CRM subsets and few-shot prompts to process semi-structured sources such as stratigraphic registers and object inventories; second, adapting open-source models via domain-specific fine-tuning and lightweight optimisation to handle the richer language of excavation narratives. The work foregrounds usability for non-technical archaeologists, rigorous validation of automatically generated RDF triples, and sustainable deployment that balances performance with energy consumption. By combining ontology guidance, human-in-the-loop review, and environmental metrics, the thesis aims to deliver a reproducible workflow that lowers the barrier to producing interoperable knowledge graphs and thus promotes data reuse and multidisciplinary inquiry in digital archaeology.

Keywords

Semantic Annotation, Large Language Models, CIDOC CRM, Knowledge Graphs, Ontology Population, Prompt Engineering

1. Introduction

The study and preservation of cultural heritage, particularly in fields like archaeology, face several challenges in the digital age. Many historical monuments and artefacts are physically fragile, geographically dispersed, or situated in locations with limited access. Consequently, researchers are turning more and more to digital resources to study, interpret, and share their findings.

Over the past decades, excavation campaigns have produced a large volume of textual documentation such as stratigraphic tables, field notes, excavation reports, and inventories of objects. While this information is rich in expert knowledge, it is often recorded in free-form narratives or heterogeneous tabular formats, making it difficult to reuse, integrate, or share in structured ways.

To address this, the cultural heritage community has promoted the use of ontologies like the CIDOC CRM [1]. It provides a formal structure to describe the entities and relationships involved in heritage documentation, enabling interoperability between institutions and digital archives. However, manually converting archaeological documentation into structured formats such as RDF that conform to CIDOC CRM is time-consuming, error-prone, and requires expertise in both the ontology and the technical tools used to apply it.

At the same time, recent advances in Natural Language Processing, particularly the emergence of LLMs, have demonstrated the capacity of machines to understand and generate structured data from natural language. Models like GPT-4, LLaMA, and Mistral have been trained on vast amounts of data and are capable of performing complex text understanding tasks with minimal examples.

This PhD research investigates whether LLMs can be applied to support the semantic annotation of archaeological documents. Specifically, it explores the use of prompt engineering and ontology-driven strategies to guide LLMs in producing CIDOC CRM-compliant RDF triples from real archaeological texts. The goal is to reduce the annotation workload for experts, increase the amount of structured

ER2025: Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos, October 20-23, 2025, Poitiers, France

ali.hariri@ensma.fr (A. Hariri)
0009-0004-9380-5138 (A. Hariri)

archaeological knowledge available, and enable more powerful forms of collaboration and data reuse across the cultural heritage sector.

The core question of the thesis is: *Can LLMs be effectively guided to produce accurate, ontology-aware semantic annotations that align with domain standards like CIDOC CRM, while remaining usable and practical for non-technical experts?* This paper outlines the current progress in addressing this question and the roadmap for the next phases of the doctoral research.

2. Background and Related Work

Ontology population has been extensively studied in various domains, including biomedicine, tourism, and cultural heritage, using a wide range of Natural Language Processing (NLP) and Machine Learning (ML) approaches [2, 3, 4]. Early work combined rule-based and statistical techniques to extract entities and relationships from domain texts. These methods offered interpretable pipelines but required substantial manual effort and expert involvement to maintain quality.

Supervised ML and Deep Learning (DL) approaches later improved scalability by learning extraction patterns from annotated corpora [5, 6]. However, these techniques often require large labeled datasets, which are rarely available in specialised domains like archaeology [7]. Constructing such datasets for heritage data is particularly difficult due to the complexity, ambiguity, and domain-specific expressions found in excavation reports or object registers.

Recent advances in transformer-based LLMs have renewed interest in automating ontology population. Pre-trained models like GPT-3, GPT-4, LLaMA, and Mistral have demonstrated promising capabilities in knowledge extraction, RDF triple generation [8], and named entity recognition [9]. In this context, cultural heritage has begun to benefit from LLM-driven annotation methods, although research remains relatively limited.

Loffredo and De Santo [10] proposed integrating ontologies with a Retrieval-Augmented Generation (RAG) framework to improve LLM performance in heritage data processing. More recently, Ding et al. [11] introduced the Knowledge Prompt Chaining approach, aiming to enhance the semantic consistency of LLM outputs by embedding graph-based constraints into prompts. Yet these studies mostly focus on prompt design and few tackle real-world annotation aligned with CIDOC CRM.

In addition to these approaches, our recent work has explored the use of ontology-subset prompting strategies for RDF extraction in the cultural heritage domain, aligned with CIDOC CRM [12]. This prior study motivates and informs the doctoral research presented in the next section.

This thesis builds on these insights by systematically evaluating LLMs for the task of ontology-guided semantic annotation using CIDOC CRM. The approach addresses both technical dimensions (ontology guidance, model adaptation) and practical constraints such as usability by non-specialists and carbon efficiency.

3. Research Objectives

This PhD research is structured in two main phases, each addressing a specific set of challenges around the use of LLMs for semantic annotation in the field of archaeology.

Phase 1: Evaluating the Extraction Capabilities of LLMs

The first phase aims to explore whether LLMs can extract structured, ontology-compliant data from semi-structured archaeological sources. These sources include tabular records such as stratigraphic unit lists or object inventories. Although more regular in format, these datasets often contain ambiguities, formatting inconsistencies, or domain-specific shorthand that complicate automatic processing.

A central challenge is the semantic classification of entities: the LLM must assign the correct CIDOC CRM class and properties to terms that may be vague or context-dependent. For example, a record might refer to a “vessel,” “pottery,” or “fragment,” all of which must be correctly understood as instances of `E22_Human-Made Object`, possibly with qualifiers about condition, function, or context. The choice of class becomes even more critical when multiple candidates are plausible (e.g., `E22` vs `E20 Biological Object`).

Another recurrent issue concerns the interpretation of formatting conventions used by archaeologists. A typical case is the notation "SU12-SU14," which appears to be a single string but actually refers to two separate stratigraphic units (SU12 and SU14). A human expert would interpret this as a range or a grouping, but an LLM without sufficient domain context may treat it as a single identifier. Similarly, punctuation, parenthetical notes, or shorthand like "?Bronze" or "Layer V?" must be handled carefully to reflect uncertainty or incomplete knowledge.

To explore these issues, we ran a series of experiments in which we varied the type of instructions given to the language model, from minimal (zero-shot) to more guided examples (one-shot, few-shot), as well as the amount of ontology context included (none, the full ontology, or a simplified subset). After each experiment, we manually reviewed the results to identify recurring errors, such as incorrect classification of concepts or confusion between related elements. This helped us understand where the model performs well and where it still needs support in the context of archaeological annotation.

Key research questions for Phase 1:

- Can LLMs accurately distinguish and apply semantically precise ontology classes in ambiguous or underspecified contexts?
- How well do LLMs interpret domain-specific text formatting, such as stratigraphic ranges, uncertain terms, or compact notations?
- What prompting strategies are most effective in improving semantic alignment and structural consistency?

An important extension of this work concerns the selection of the ontology subset. In the current phase, the choice of classes and properties is carried out manually, based on expert knowledge of both the archaeological domain and the CIDOC CRM ontology. While this ensures high-quality alignment, it also introduces a significant dependency on human expertise and limits scalability to other datasets or domains.

Future developments will therefore explore automatic methods for identifying the most relevant portions of the ontology. The idea is to compute semantic proximity between the data to be annotated and the ontology vocabulary, so that classes and properties most closely related to the observed content are prioritized. By combining these techniques , it may become possible to highlight candidate classes and properties automatically, leaving the experts with the lighter task of validation rather than full manual selection.

Such an approach would not only reduce the manual effort required from domain specialists but also provide a more adaptive mechanism, capable of tailoring the ontology subset dynamically to the specific characteristics of each dataset. For instance, datasets focusing on archaeological objects might emphasize material-related properties, while those centered on excavation activities could highlight temporal and spatial relations. This flexibility would help strike a better balance between semantic precision and coverage, while maintaining the trustworthiness of the resulting annotations.

Phase 2: Sustainable Deployment, Querying and Result Integration

The second phase places emphasis on the sustainable and practical deployment of ontology-based annotation workflows. While large language models such as GPT-4 demonstrate strong performance, they also raise concerns due to their high computational requirements, significant energy consumption, and reliance on cloud-based infrastructures that may be inaccessible to many archaeological institutions.

To address these challenges, this phase explores strategies that enable more resource-efficient deployment without compromising annotation quality. Possible directions include domain-specific fine-tuning of smaller open-source models, quantization techniques to reduce memory and processing demands, and hybrid workflows in which uncertain cases are flagged for human validation rather than fully automated processing. In addition, carbon monitoring tools will be integrated into the workflow to make the environmental impact of computational choices transparent and measurable.

Beyond efficiency, another central objective concerns the usability of the produced annotations. The ultimate goal of the querying process is to enable researchers to address complex domain-specific questions that cannot be answered by a single document in isolation. By exploiting the semantic structure of the annotations, it becomes possible to perform sophisticated queries that combine evidence from multiple sources.

Key research questions for Phase 2:

- To what extent can domain-specific fine-tuning of language models improve alignment with established ontologies such as CIDOC CRM?
- What are the trade-offs between model size, annotation quality, and energy consumption?

4. Case Study and Corpus

This research is grounded in the study of the *Hypogeum of the Dunes*, a Merovingian funerary complex located in Poitiers, France. The site is under investigation as part of the French national research project ANR Digitalis, which seeks to develop digital methods and tools for multidisciplinary archaeological analysis. The hypogeum offers a rich case for exploring semantic annotation due to its stratigraphic complexity, long excavation history, and interdisciplinary documentation.

This study uses a combination of structured and unstructured archaeological documents to evaluate the performance of LLM-based semantic annotation. On the one hand, the corpus includes structured datasets, such as stratigraphic unit registers, inventories, and excavation tables, extracted from expert-produced documents and transformed into consistent CSV formats. On the other hand, it incorporates unstructured textual content, including field notes, narrative observations, academic articles, and monographs written by archaeologists over several campaigns.

Among the structured datasets used in our experiments, four stand out as representative examples of archaeological documentation commonly encountered in fieldwork and post-excavation analysis:

- *Stratigraphic Unit Register* (182 rows, 8 columns): Describes stratified deposits, including unit identifiers, layer types, relationships, and materials.
- *Lapidary Inventory* (68 rows, 12 columns): Lists movable stone artefacts, with dimensions, typological classifications, inscriptions, and provenance data.
- *Geological Sample Table* (68 rows, 9 columns): Documents samples of building stones and geological sources, with mineralogical and locational information.
- *Excavation Facts Table* (80 rows, 10 columns): Summarises observed events and operations, including actors, dates, and stratigraphic relations.

These datasets reflect real-world heterogeneity: abbreviations, typographical noise, compact notations (e.g., `SU12–SU14`), and inconsistent delimiters are common. Some tables use nested content (e.g., multiple values in a single cell), while others embed uncertainty indicators (e.g., “possibly”, “?”) that require careful modelling.

To prepare data for LLM input, each dataset was split into 20-row segments, forming the textual base of individual prompts. This chunking ensures compatibility with model context limitations and preserves coherence within each block. The RDF output from each chunk is evaluated manually and compared against expert expectations.

This corpus provides a unique opportunity to assess LLM performance in a real and messy archaeological context offering insights into semantic interpretation, ontology alignment, and practical deployment.

5. Methodology

The methodological approach is based on two complementary strategies that define the two main contributions of this work:

Phase 1: Ontology-guided Prompt Engineering

The first phase seeks to evaluate to what extent large language models can generate RDF triples compliant with CIDOC CRM from semi-structured archaeological sources. To this end, we designed a methodology based on prompt engineering that combines different forms of ontological guidance with varying levels of example provision (Figure 1). The idea is to evaluate how different forms of ontological context influence the model's performance.

Three degrees of ontological context were explored. In the first case, the model was left unguided, relying only on its pre-training, which already includes references to CIDOC CRM. In the second, the complete ontology was included in the prompt, which provided stronger guidance but also resulted in very long prompts. Finally, a third approach used a curated subset tailored to archaeology, prioritizing the most relevant classes and properties. This approach reduces hallucinations while directing the model toward more precise results.

In parallel, we examined the effect of different prompting strategies. With zero-shot prompting, the model must rely solely on textual instructions. With one-shot prompting, a single annotated example provides a transformation template from text to RDF. Few-shot prompting extends this by including several varied examples drawn from the corpus, which strengthens the model's ability to generalize across the heterogeneity of archaeological data.

This strategy is primarily applied to structured and semi-structured datasets, where the document layout already provides a partial indication of the semantic content. The prompts are designed to translate this structure into explicit semantic annotations by helping the model reason within the CIDOC CRM framework. This first step, illustrated in the left-hand part of Figure 1, establishes a foundational understanding of how general-purpose LLMs interpret archaeological data and apply ontological logic.

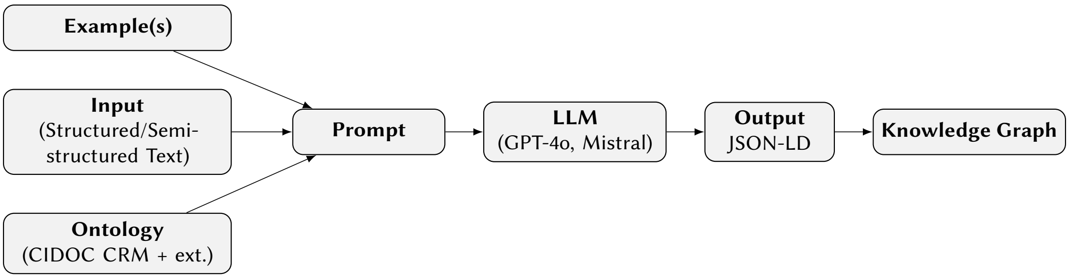
Example(s)
Input (Structured/Semi-structured Text)
Ontology (CIDOC CRM + ext.)
Prompt
LLM (GPT-4o, Mistral)
Output JSON-LD
Knowledge Graph

Figure 1: General workflow: semantic annotation using LLMs guided by ontology and examples.

Phase 2: Sustainable Deployment, Querying and Result Integration

The methodological direction of this work is oriented towards adapting large language models to the specific requirements of ontology-driven knowledge extraction in archaeology. Current state-of-the-art LLMs, although pre-trained on large amounts of open web data, including CIDOC CRM description, remain generic-purpose systems. They are not designed for structured extraction tasks and often exhibit shortcomings such as reduced accuracy in ontology population and susceptibility to hallucinations. For this reason, we propose the development of a domain-adaptive strategy that specifically addresses these issues.

The approach will focus on the fine-tuning of a smaller open-source language model, rather than the use of large proprietary systems. This choice is motivated both by sustainability concerns, smaller models require fewer computational resources and generate lower carbon emissions, and by the need to improve task-specific performance through specialization.

To achieve this, the fine-tuning procedure will rely on a dual-source training corpus. On one side, domain-relevant archaeological descriptions from open-access datasets will provide empirical grounding in the language of archaeological documentation. On the other side, the complete textual documentation of CIDOC CRM (including classes, properties, scope notes, and hierarchical relations) will supply explicit ontological knowledge. This combination is expected to allow the model to acquire both contextual and semantic grounding, improving its ability to align extracted entities and relations with CIDOC CRM concepts.

The specialized model will be designed to address two core tasks: (i) entity recognition, for the identification of archaeological entities and their alignment with CIDOC CRM classes, and (ii) relation extraction, for the detection of semantic links and their mapping to CIDOC CRM properties. By integrating both empirical texts and ontological documentation, the model is expected to reduce hallucinations, enhance semantic consistency, and improve the overall quality of ontology-based annotations.

Evaluation Strategy

To evaluate the performance of the proposed approach, we use three main metrics:

- *Precision and Recall*: Computed at the RDF triple level by comparing LLM outputs with gold-standard annotations validated by domain experts.
- *Competency Question (CQ) Score*: Measures the ability of the generated knowledge graph to correctly answer domain-specific SPARQL queries.
- *Carbon Footprint*: The energy consumption of each model is tracked using CodeCarbon and EcoLogits, allowing us to quantify CO_2 emissions per annotation task.

The application of these two phases allows us to progressively address the complexity of the task. In the first phase, we focus on understanding how LLMs behave when exposed to domain-specific prompts and structured data, and we evaluate their capability to extract semantically accurate knowledge aligned with ontologies such as CIDOC CRM. This phase provides insight into the conditions under which LLMs succeed or fail in knowledge extraction.

In the second phase, we aim to go beyond experimentation and design a robust, sustainable system. By fine-tuning models on archaeological texts and controlling model size and inference cost, we move toward creating specialised expert models that are both high-performing and more eco-efficient. This positions the work not only as a technical investigation but also as a contribution to sustainable digital practices in heritage data management.

6. Discussion and Open Questions

This doctoral research raises several open questions that touch on technical, methodological, and practical dimensions of LLM-assisted semantic annotation.

Semantic Precision vs. Ontological Coverage. According to our initial results, the use of curated ontology subsets has proven to be the most effective strategy for guiding LLMs to generate precise, semantically aligned RDF annotations. This setup helps limit ambiguity, simplifies the classification space, and improves the reliability of extracted triples. However, it may also omit low-frequency concepts or rare relationships that are important for more advanced or exploratory queries. Finding the right balance between simplicity and completeness remains a key challenge. One possible direction is to develop adaptive subset selection mechanisms that evolve dynamically based on the user's research needs or the semantic richness of the input corpus.

Structured vs. Unstructured Input. The performance of prompt-based LLMs varies depending on the structure and regularity of the input. While structured tables can be more easily aligned with RDF schemas, unstructured narratives are harder to process but often contain richer insights. Methods that can adapt dynamically to input heterogeneity are still underexplored.

From Model Output to Trustworthy Knowledge. One of the key concerns when using LLMs for semantic annotation in archaeology is how to ensure the reliability and interpretability of their outputs. While these models can generate RDF triples that appear structurally correct, their reasoning process remains opaque. This is problematic in a domain like heritage, where expert trust and data traceability are essential.

To make LLM-generated annotations trustworthy, we consider integrating multiple layers of validation. These include confidence scores, links to source sentences, and explicit rationale for class/property selection. Human-in-the-loop strategies are also essential, especially for ambiguous cases, allowing experts to review and confirm machine-generated suggestions.

Sustainability Trade-offs. The environmental impact of using state-of-the-art LLMs is significant, especially when relying on large commercial models. This raises questions about the viability of widespread deployment in low-infrastructure or resource-conscious contexts. Many archaeological institutions lack access to powerful GPUs or cloud computing services, making it impractical to rely on high-end models like GPT-4 on a daily basis. To address this, we are exploring lighter models (such as Mistral or LLaMA), as well as model quantization techniques that reduce energy consumption while maintaining acceptable performance. Another promising avenue is to combine local inference for standard cases with human review for ambiguous ones. This hybrid strategy may provide a balance between automation, interpretability, and ecological responsibility.

Generalisation and Transferability. A key question that emerges from our study is how well models trained or fine-tuned on one dataset, such as the Hypogeum of the Dunes, can generalise to other archaeological corpora or related domains like epigraphy, conservation, or museum cataloguing. In practice, different heritage domains often use distinct terminologies, notational conventions, and descriptive norms, which may reduce the model's ability to transfer learned patterns. This makes it necessary to explore strategies such as transfer learning, few-shot adaptation, or domain-invariant feature extraction. Understanding these limitations is essential for designing annotation tools that are truly reusable across institutions and use cases.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5 in order to: Grammar and spelling check.

After using these service, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Doerr, The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata, AI magazine 24 (2003) 75–75.

[2] D. Maynard, Y. Li, W. Peters, Nlp techniques for term extraction and ontology population., 2008.

[3] C. Biemann, Ontology learning from text: A survey of methods, Journal for Language Technology and Computational Linguistics 20 (2005) 75–93.

[4] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, E. Zavitsanos, Ontology population and enrichment: State of the art, Knowledge-driven multimedia information extraction and ontology evolution: Bridging the semantic gap (2011) 134–166.

[5] A. Imsombut, C. Sirikayon, An alternative technique for populating thai tourism ontology from texts based on machine learning, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science, IEEE, 2016, pp. 1–4.

[6] P. Sambandam, D. Yuvaraj, P. Padmakumari, S. Swaminathan, Spiking equilibrium convolutional neural network for spatial urban ontology, Neural Processing Letters 55 (2023) 7583–7602.

[7] F. Suchanek, G. Ifrim, G. Weikum, Leila: Learning to extract information by linguistic analysis, in: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 2006, pp. 18–25.

[8] R. R. de Souza, T. L. Pinheiro, J. C. B. Oliveira, J. C. dos Reis, Knowledge graphs extracted from medical appointment transcriptions: Results generating triples relying on llms., in: KEOD, 2023, pp. 129–139.

[9] M. Freund, R. Dorsch, S. Schmid, T. Wehr, A. Harth, Enriching rdf data with llm based named entity recognition and linking on embedded natural language annotations, in: International Knowledge Graph and Semantic Web Conference, Springer, 2024, pp. 109–122.

[10] R. Loffredo, M. De Santo, Using ontologies for llm applications in cultural heritage (2024).

[11] N. P. Ding, J. Du, Z. Feng, Knowledge prompt chaining for semantic modeling, arXiv preprint arXiv:2501.08540 (2025).

[12] A. Hariri, S. Jean, M. Baron, Towards automating rdf extraction for archaeological knowledge graphs with llms, in: International Conference on Database and Expert Systems Applications, Springer, 2025, pp. 83–97.