AI assisted extraction of archaeological knowledge from legacy records

Nevio Dubbini (University of Pisa), *nevio.dubbini@unipi.it*
Quirino Saraceni (University of Pisa), *quirino.saraceni@phd.unipi.it*
Gabriele Gattiglia (University of Pisa), *gabriele.gattiglia@unipi.it*
Francesco D'Antoni (Sapienza University of Rome), *francesco.dantoni@uniroma1.it*

Keywords
Information extraction; Human in the loop; Archaeological legacy data; Artificial intelligence; Large Language Models

The archaeological record has been extensively documented since the late 19th century, producing an immense corpus of reports, notebooks, correspondence, and administrative forms. These documents constitute an important source for reconstructing historical phenomena through archaeological evidence. However, much of this documentation, i.e. the so-called legacy data, remains locked in analogue formats or in unstructured digital "grey literature", using file formats from which data cannot be easily extracted, such as PDF and JPG. The lack of structured, machine-readable data significantly limits the possibility of accessing, comparing, and computationally analysing this historical information at scale. Manual digitisation and information extraction from these sources are time-consuming and expensive, requiring expert interpretation and domain-specific knowledge. In recent years, Artificial Intelligence (AI) and Natural Language Processing (NLP) have shown great promise in automating parts of these processes, particularly through Optical Character Recognition (OCR), Named Entity Recognition (NER), and, more recently, transformer-based Large Language Models (LLMs). Yet, the heterogeneous nature of archaeological documentation, ranging from digital-born reports to typewritten or handwritten excavation notes, poses substantial challenges for automated systems.

This work addresses the problem of extracting structured archaeological and historical information from legacy documents by developing and evaluating an information-extraction workflow that combines AI-based tools with expert-driven validation. The study focuses on defining a robust methodology for automatically identifying and structuring key entities from unstructured texts while ensuring data reliability through a human-in-the-loop validation process. The work draws upon the MAGOH database (Managing Archaeological data for a sustainable Governance of the Heritage) as a test environment, encompassing over 10,000 excavation records from the provinces of Pisa, Florence, Prato, Pistoia, and Lucca, Italy.

The workflow integrates document digitisation via OCR, information extraction and summarisation using NLP and LLMs, and human-in-the-loop validation. The dataset includes both digital-born and scanned PDF documents, which are converted into machine-readable text using OCR methods. Documents are subsequently segmented and ranked using NLP techniques, including Named Entity Recognition, keyword extraction, sentence-embedding similarity, and relevance scoring. The most contextually relevant sections are submitted to an LLM through structured prompts designed to identify key archaeological entities, including site name, stratigraphic unit, material type, chronological attribution, and geographic reference. These categories are particularly relevant for transforming individual archaeological documents into structured evidence that can be queried and compared across sites, periods, and territories. Recognising the semantic ambiguity and contextual variability of archaeological texts, the workflow integrates a human-in-the-loop validation stage. For each extracted piece of information, the system highlights the exact text lines in the original document from which the data were derived, enabling specialists to inspect and verify the

extraction context. Each output is also accompanied by a confidence score, allowing uncertain or ambiguous results to be prioritised for expert review.

Initial results demonstrate the potential of hybrid AI-human workflows to accelerate the structuring of archaeological legacy data. OCR and preprocessing achieved conversion accuracy above 95% for typewritten and image documents, while the LLM-based entity-recognition process achieved an F1 score of approximately 0.82. Expert validation substantially reduced the time required compared with traditional manual data entry, while maintaining the traceability of extracted information.

Beyond the digitisation of archaeological archives, this approach contributes to the development of computational methods for historical research. Transforming heterogeneous legacy documentation into structured, machine-readable information enables the aggregation of archaeological evidence across large documentary corpora and facilitates its subsequent use for investigating spatial, chronological, and material patterns in the past. The study therefore demonstrates how AI and LLM-based information extraction, combined with expert oversight, can provide a scalable and reproducible method for transforming archaeological legacy records into structured historical evidence while preserving the contextual and interpretative rigour required by historical and archaeological research.

REFERENCES

Anichini, F., Gattiglia G., and Saponara A.R.. 2023. "MAGOH: un nuovo strumento per la gestione e la consultazione dei dati archeologici del Nord della Toscana". Archeologia e Calcolatori 34 (2): 277–96. https://doi.org/10.19282/ac.34.2.2023.15.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T.T., Moazam, H., Miller, H., Zaharia, M. and Potts C. 2024. "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines". The Twelfth International Conference on Learning Representations.

Seow, W.L., Chaturvedi, I., Hogarth, A., Mao, R. and Cambria E. 2025. "A Review of Named Entity Recognition: From Learning Methods to Modelling Paradigms and Tasks". Artificial Intelligence Review 58 (10): 315. https://doi.org/10.1007/s10462-025-11321-8.