YUGARTH - Autonomous Archaeological Discovery Engine

Vaishnavi Praveen Rathi[1] Pranavi Rajesh Bhosale[2] Yash Vilas Chorghe[3] Tanushka Amrut Herekar[4]

[1,2,3,4]Department of Computer Science and Technology

[1,2,3,4]Zeal Polytechnic, Pune, Maharashtra, India

Abstract — Archaeological information is distributed across inscriptions, excavation reports, artifact catalogues, and historical publications. Connecting these records is challenging when names vary, dates are uncertain, and evidence is incomplete. This survey reviews artificial intelligence techniques for archaeological knowledge discovery, including natural language processing, knowledge graphs, relationship prediction, and evidence retrieval. It also introduces YUGARTH, a proposed framework for exploring historical connections, retrieving supporting evidence, and checking chronological compatibility. The framework aims to assist researchers in investigating potential relationships rather than automatically establishing historical facts.

Keywords: Archaeology, Artificial Intelligence, Natural Language Processing, Knowledge Graphs, Historical Data, Evidence Retrieval

I. Introduction

Archaeological research involves studying material remains and documentary evidence to understand past societies. Important information about historical people, places, artifacts, and events may be distributed across multiple publications and digital collections. Conventional keyword searches can retrieve individual documents but may not reveal indirect relationships between records.

Artificial intelligence can support this process by extracting information from text, organizing entities into knowledge graphs, and identifying potential connections for further investigation. Previous research has explored NLP for archaeological document access and knowledge graphs for cultural-heritage information management [1], [2], [5].

YUGARTH proposes an evidence-oriented archaeological discovery framework. Its purpose is to help researchers investigate relationships among historical records, rather than automatically declare an unverified connection to be a historical fact.

II. Literature Review

A. NLP for Archaeological Documents

Natural language processing enables computers to process and organize information contained in written documents. In archaeology, NLP techniques can improve access to archaeological grey literature and help researchers locate relevant information in large collections [1]. However, document retrieval alone does not necessarily connect entities mentioned across separate sources.

B. Knowledge Graphs for Historical Relationships

A knowledge graph represents information through entities and relationships. For example, a historical site, an artifact, and a person can be represented as connected nodes when reliable evidence establishes their relationships. Graph-based exploration can help researchers investigate patterns that are difficult to identify through isolated document searches [2].

C. AI-Assisted Cultural-Heritage Research

Research has explored using generative language models to construct knowledge graphs from scholarship on antiquities trafficking [3]. Other work has combined knowledge graphs and deep learning to extract and connect cultural-heritage information, including a case study involving ceramic records from the Palace Museum in China [4].

These approaches demonstrate relevant methods, but they address different tasks and datasets. Challenges include incomplete records, ambiguous entity names, extraction errors, missing relationships, and the need to verify AI-generated connections against original sources.

III. Comparative Analysis and Research Gaps

Approach	Main contribution	Limitation
NLP-based indexing [1]	Improves document discovery	May not connect indirect relationships
Knowledge-graph exploration [2]	Represents connected information	Depends on data quality
Language-model-assisted graph construction [3]	Extracts candidate relationships from text	Requires verification
Deep-learning-based heritage graphs [4]	Supports entity extraction and graph completion	Depends on the domain and dataset

Table I: Comparison of Related Approaches

The reviewed approaches address document retrieval, relationship representation, and knowledge extraction. A potential research direction is to integrate these capabilities with source-passage retrieval and chronological consistency checks in a single archaeological research workflow. YUGARTH proposes this integration as a framework to investigate. Its contribution must be evaluated through implementation and comparison with existing approaches; novelty and performance cannot be assumed from the design alone.

IV. Proposed YUGARTH Framework

YUGARTH is proposed as a research-support system that connects information extracted from archaeological documents and presents candidate relationships with their supporting evidence

The framework uses natural language processing to identify entities such as historical people, locations, artifacts, and dates, then organizes them into a connected knowledge graph.

All rights reserved by www.ijsrd.com

It also aims to rank potential relationships, retrieve relevant source passages, and flag possible chronological inconsistencies.

The final interpretation remains subject to verification by archaeological researchers.

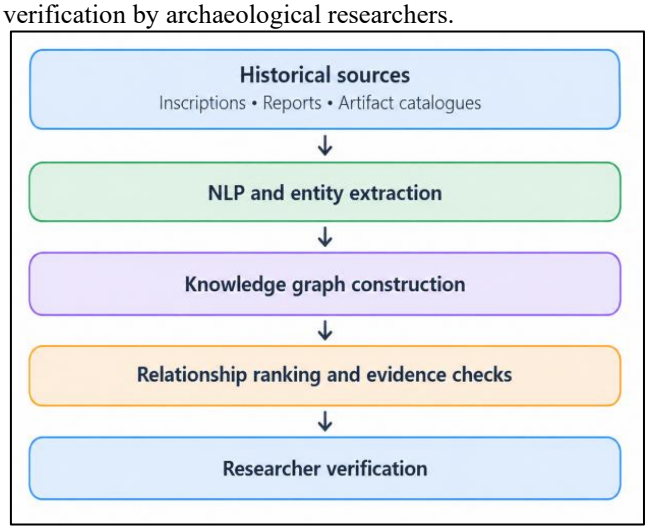


Fig. 1: Proposed workflow of the YUGARTH archaeological knowledge discovery framework.

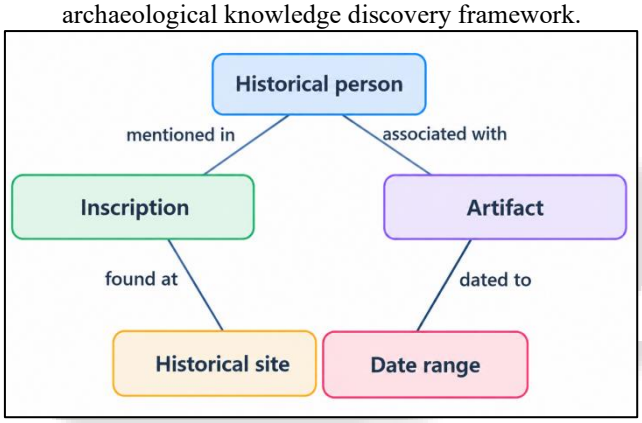


Fig. 2: Conceptual example of entities and relationships in an archaeological knowledge graph; the links are illustrative, not verified historical claims.

A. Data Collection and Processing

The system would use permitted, documented archaeological sources such as digitized publications, catalogues, and excavation reports.

Text preprocessing would standardize formatting and identify candidate entities, including people, locations, artifacts, and dates. Uncertain or damaged text should be marked as uncertain instead of silently reconstructed.

B. Knowledge Graph Construction

Extracted entities would become graph nodes, while documented relationships would become edges. Each relationship should retain source information, such as a document identifier and the passage from which it was extracted. This would allow users to inspect the evidence behind a connection.

C. Relationship Ranking and Evidence Verification

YUGARTH could rank candidate relationships using textual evidence, graph structure, and chronological compatibility. One possible proposed scoring formulation is:

$$S(r) = \alpha E(r) + \beta G(r) + \gamma T(r) \quad (1)$$

where:

S(r) represents the score of candidate relationship r.
E(r) represents the strength of its textual evidence.
G(r) represents its graph-based relationship score.
T(r) represents its chronological compatibility score.
α, β, and γ represent non-negative weighting parameters.

If all component scores are normalized between 0 and 1, the weights can be constrained so that $\alpha + \beta + \gamma = 1$. The component scores and weights would need to be defined and tested during implementation. Equation (1) is illustrative, not a validated formula or a reported result from an existing study.

The system should retrieve source passages that support or challenge each candidate connection. Chronological checks could flag relationships that conflict with recorded date ranges. A high score should indicate a candidate worth investigating, not proof that the relationship is historically correct.

D. Evaluation Plan

A prototype could be evaluated on a curated dataset containing source documents and expert-reviewed entity relationships. Entity extraction could be measured using precision, recall, and F1-score.

Relationship ranking could be evaluated using Mean Reciprocal Rank (MRR) or Hits@K, while evidence retrieval could use Precision@K. The evaluation should compare YUGARTH with suitable baseline methods and report measured results only after experiments are conducted.

V. Conclusion

This survey reviewed AI techniques relevant to archaeological information discovery, including NLP, knowledge graphs, and language-model-assisted extraction. YUGARTH proposes combining these techniques with evidence retrieval and chronological checks to help researchers explore possible relationships across historical records. Its evidence-oriented design aims to make candidate connections easier to inspect and verify. Future work should implement the framework, evaluate it on a suitable archaeological dataset, and examine extraction errors, incomplete evidence, and performance against baseline approaches. Human expertise remains essential when interpreting uncertain historical information.

References

[1] A. Vlachidis, C. Binding, D. Tudhope, and M. May, "Excavating grey literature: A case study on the rich indexing of archaeological documents via natural language-processing techniques and knowledge-based resources," Aslib Proceedings, vol. 62, nos. 4–5, pp. 466–475, 2010, doi: 10.1108/00012531011074708.

[2] Y. Y. Huang, S. S. Yu, J. J. Chu, H. Fan, and B. Du, "Using knowledge graphs and deep learning algorithms to enhance digital cultural heritage management," npj Heritage Science, vol. 11, art. 204, 2023, doi: 10.1186/s40494-023-01042-y.

[3] G. Gattiglia, "Managing Artificial Intelligence in Archaeology: An Overview," Journal of Cultural

All rights reserved by www.ijsrd.com

Heritage, vol. 71, pp. 225–233, 2025, doi: 10.1016/j.culher.2024.11.020.

[4] C. Binding, D. Tudhope, and A. Vlachidis, “A study of semantic integration across archaeological data and reports in different languages,” Journal of Information Science, vol. 45, no. 3, pp. 364–386, 2019, doi: 10.1177/0165551518789874.

All rights reserved by www.ijsrd.com