Adentrándonos en el conocimiento del patrimonio arqueológico a través del RAG de IDEArq

Tema: Nº. 4 - Inteligência artificial

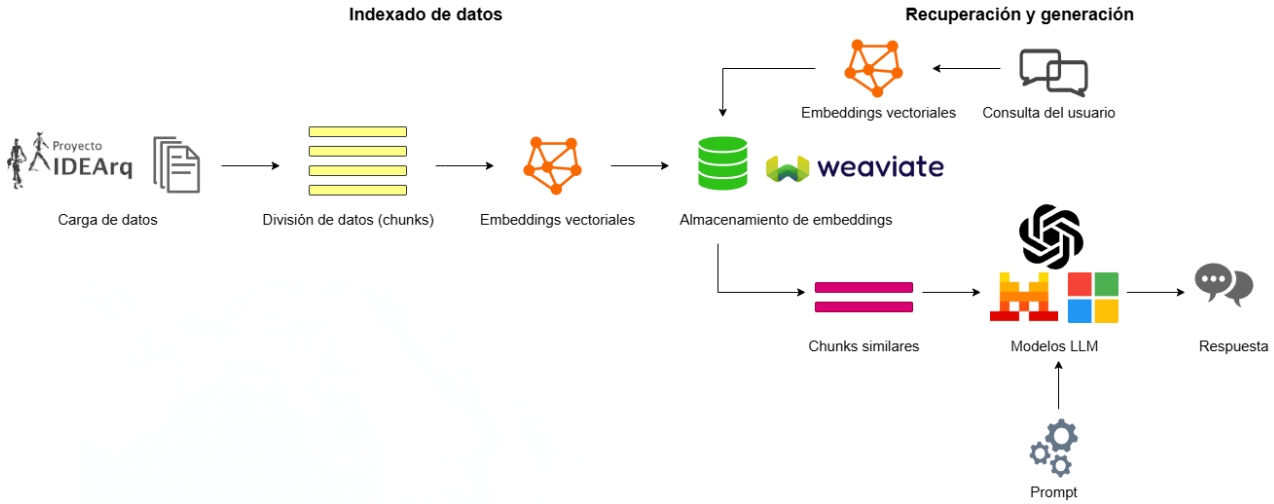
Resumo

La plataforma EXEA, desarrollada en el Centro de Ciencias Humanas y Sociales (CCHS) del Consejo Superior de Investigaciones Científicas (CSIC), continúa consolidándose como infraestructura digital de referencia para la gestión, preservación e interoperabilidad del conocimiento científico en el ámbito de las Ciencias Sociales y Humanidades. En el marco de esta plataforma, **IDEArq —Infraestructura de Datos Espaciales especializada en Arqueología** del CSIC— ha experimentado avances en su integración con técnicas de Inteligencia Artificial, ampliando las capacidades del sistema RAG (*Retrieval-Augmented Generation*)[1] presentado en la edición anterior.

El sistema RAG, que combina grandes modelos de lenguaje (LLM) con mecanismos de recuperación semántica de información, ha incorporado tres mejoras en su arquitectura. En primer lugar, se ha implementado una estrategia de búsqueda híbrida que combina recuperación densa (mediante *embeddings* vectoriales) con recuperación dispersa (basada en palabras clave), logrando una cobertura más completa y robusta del corpus documental. En segundo lugar, se ha integrado un módulo de *multiquery*, que genera automáticamente variantes de la consulta original para ampliar el espacio de búsqueda y reducir la dependencia de formulaciones específicas por parte del usuario. Por último, se ha incorporado un mecanismo de *reranking* que reordena los fragmentos recuperados en función de su relevancia semántica respecto a la consulta, mejorando la calidad del contexto que se proporciona al modelo generativo y, con ello, la precisión de las respuestas obtenidas.

Paralelamente, se ha desarrollado un grafo bibliográfico de **IDEArq** que estructura de forma explícita las relaciones entre publicaciones científicas y autores, integrando este recurso en la arquitectura del sistema RAG para enriquecer las respuestas a consultas de carácter bibliográfico. Este grafo permite no solo recuperar referencias relevantes, sino también contextualizar la producción científica asociada a cada conjunto de datos, facilitando la trazabilidad del conocimiento y la identificación de líneas de investigación.

En conjunto, estos avances consolidan la evolución de IDEArq hacia una plataforma de conocimiento abierto que integra datos espaciales, literatura científica y representación formal del conocimiento, reforzando su capacidad para apoyar la investigación interdisciplinar. El proyecto, desarrollado en colaboración entre la Universidad Politécnica de Madrid y la Unidad SIG y Humanidades Digitales junto al Instituto de Historia del CCHS-CSIC, sienta las bases para extender este enfoque a otras disciplinas dentro del ecosistema EXEA.



Referencias y vínculos:

- [1] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Recursos adicionales

- EXEA Platform: https://exea.csic.es/es
- IDEArq: http://www.idearqueologia.org/
- CSIC: https://www.csic.es/es
- CCHS-CSIC: https://www.cchs.csic.es/es
- UPM: https://www.upm.es/
- Repositorio GitHub: https://github.com/momentumSIG/IDEArqRAG

Palavras chave

Arqueología, *Retrieval-Augmented Generation* (RAG), LLM, IDE

Autores

Elena Aguayo Jara
elena.aguayo@cchs.csic.es
Unidad SIGyHD (CCHS-CSIC)

Nuria Hermida Jiménez
nuria.hermida@cchs.csic.es
Unidad SIGyHD (CCHS-CSIC)

Juan M. Vicent García
juan.vicent@cchs.csic.es
Instituto de Historia (IH-CCHS-CSIC)

Luis M. Vilches-Blázquez
luis.vilches@upm.es
ETSI en Topografía, Geodesia y Cartografía
Universidad Politécnica de Madrid