

Explorando el conocimiento arqueológico con IA: Un sistema RAG sobre recursos de IDEArq

Elena Aguayo Jara¹, Nuria Hermida Jiménez¹, Juan M. Vicent García², Luis M. Vilches-Blázquez³

Unidad de Sistemas de Información Geográfica y Humanidades Digitales (uSIGyHD)¹

Instituto de Historia (IH)²

Universidad Politécnica de Madrid (UPM)³

Centro de Ciencias Humanas y Sociales (CCHS-CSIC)



EXEA - Transformación digital del patrimonio arqueológico

EXEA constituye una infraestructura digital para la gestión y difusión del patrimonio arqueológico. Desarrollada por el Centro de Ciencias Humanas y Sociales del CSIC.

<https://exea.csic.es/es>

exea

Acerca de Datos y Servicios Geoportal IMAGO ORBIS Corpora Textuales Recursos SCyHD

Centros Humanos y Sociales del CSIC

Apoyo a la investigación, procesamiento de datos, acceso, visualización, interoperabilidad, innovación

Infraestructura de Datos Espaciales (IDE) IMAGO ORBIS

Accede a datos arqueométricos y Arte Rupestre de yacimientos de la Península Ibérica

IDE-Arq: Infraestructura de Datos Espaciales de Investigación Arqueológica

Publicación de datos científicos arqueológicos georreferenciados de acceso abierto.

IDEArq es una plataforma centrada en la difusión de datos temáticos de investigación arqueológica. Cuenta con **casi 2000 yacimientos arqueológicos de la Península Ibérica**, permitiendo el acceso mediante servicios WMS, visualizador cartográfico y consulta directa en su plataforma online.



Acceso a IDEArq:

<http://www.idearqueologia.org>



Geoservicios y Metadatos



Visualizador cartográfico



Base de Datos Carbono 14



contacto: idearq@cchs.csic.es

Grupo de Investigación Prehistoria Social y Económica del Instituto de Historia (IH).

Laboratorio de Arqueología del Paisaje y Teledetección (LabTel) del Instituto de Historia (IH).

Unidad de Sistemas de Información Geográfica y Humanidades Digitales (uSIGyHD) del CCHS.



Carbono 14

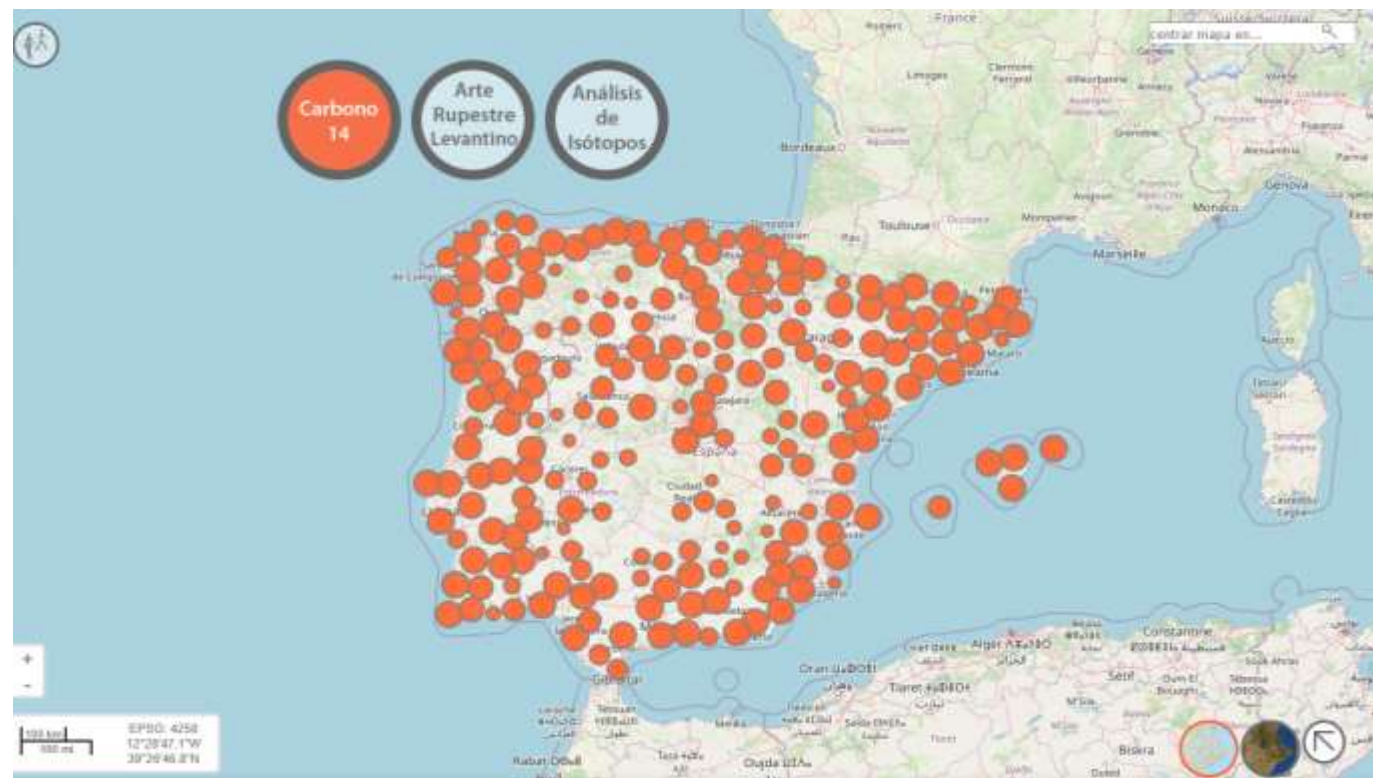
Pintura
rupestre
levantina

Análisis de
isótopos



IDEArq – Dataciones de C¹⁴

- Península Ibérica e Islas Baleares.
- **Epipaleolítico/Mesolítico hasta la Edad del Hierro.**
- **13.683** dataciones.



Carondio

Villayón (Asturias, ES)

Descriptor:

Funerario, Necrópolis (Edad del Bronce)

Info

<<

>>

Seleccionar

Carondio

Yacimiento

Dataciones C14

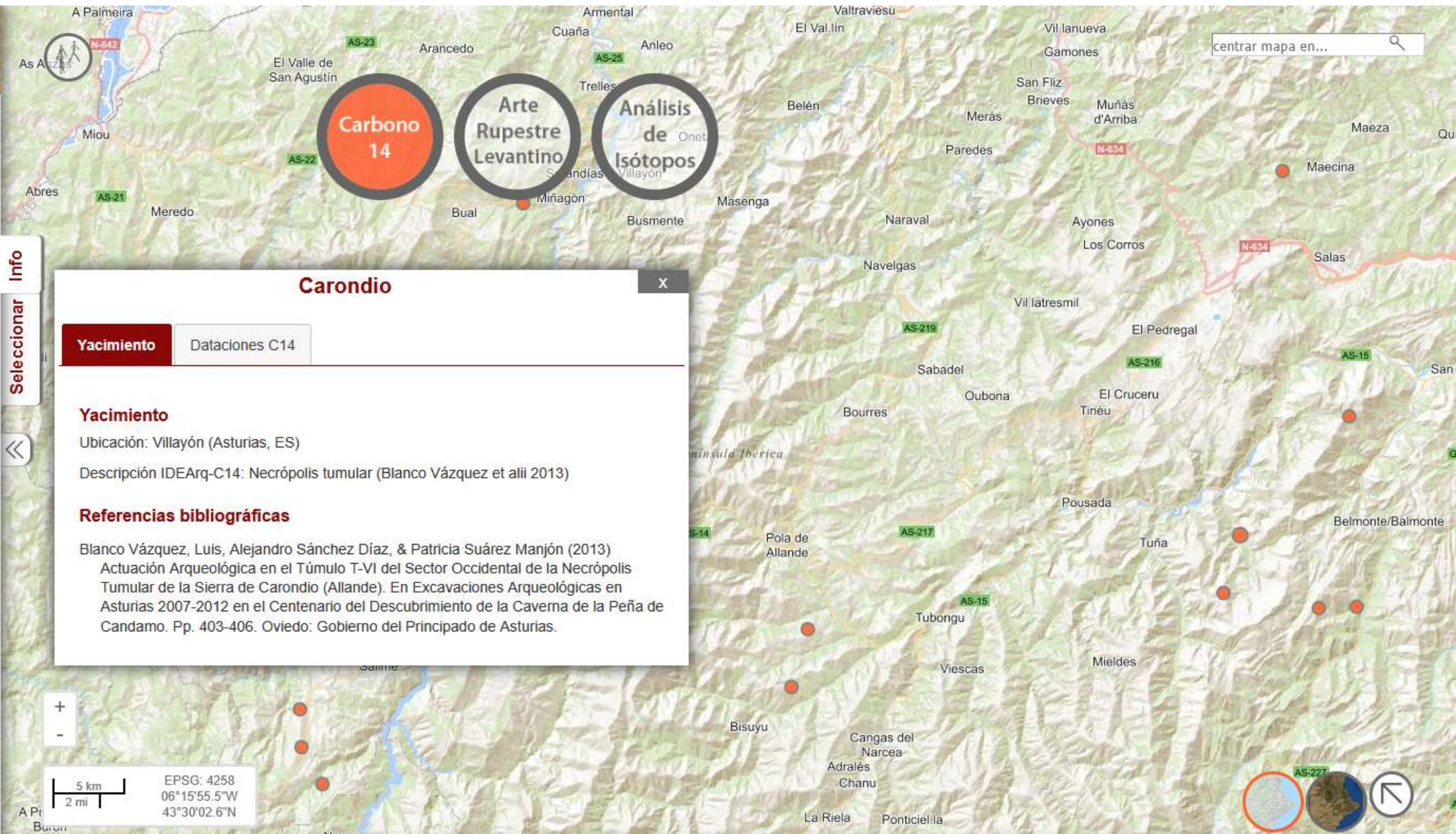
Yacimiento

Ubicación: Villayón (Asturias, ES)

Descripción IDEARQ-C14: Necrópolis tumular (Blanco Vázquez et alii 2013)

Referencias bibliográficas

Blanco Vázquez, Luis, Alejandro Sánchez Díaz, & Patricia Suárez Manjón (2013) Actuación Arqueológica en el Túmulo T-VI del Sector Occidental de la Necrópolis Tumular de la Sierra de Carondio (Allande). En Excavaciones Arqueológicas en Asturias 2007-2012 en el Centenario del Descubrimiento de la Caverna de la Peña de Candamo. Pp. 403-406. Oviedo: Gobierno del Principado de Asturias.



Español | English

Información

Busca un yacimiento

Carondio

Villayón (Asturias, ES)

Descriptor:

Funerario, Necrópolis (Edad del Bronce)

Info

Seleccionar



Carondio

X

Yacimiento

Dataciones C14

Dataciones C14

Beta-324748 **3110 ± 30 BP** sobre Sedimento (AMS) +info

Referencias bibliográficas

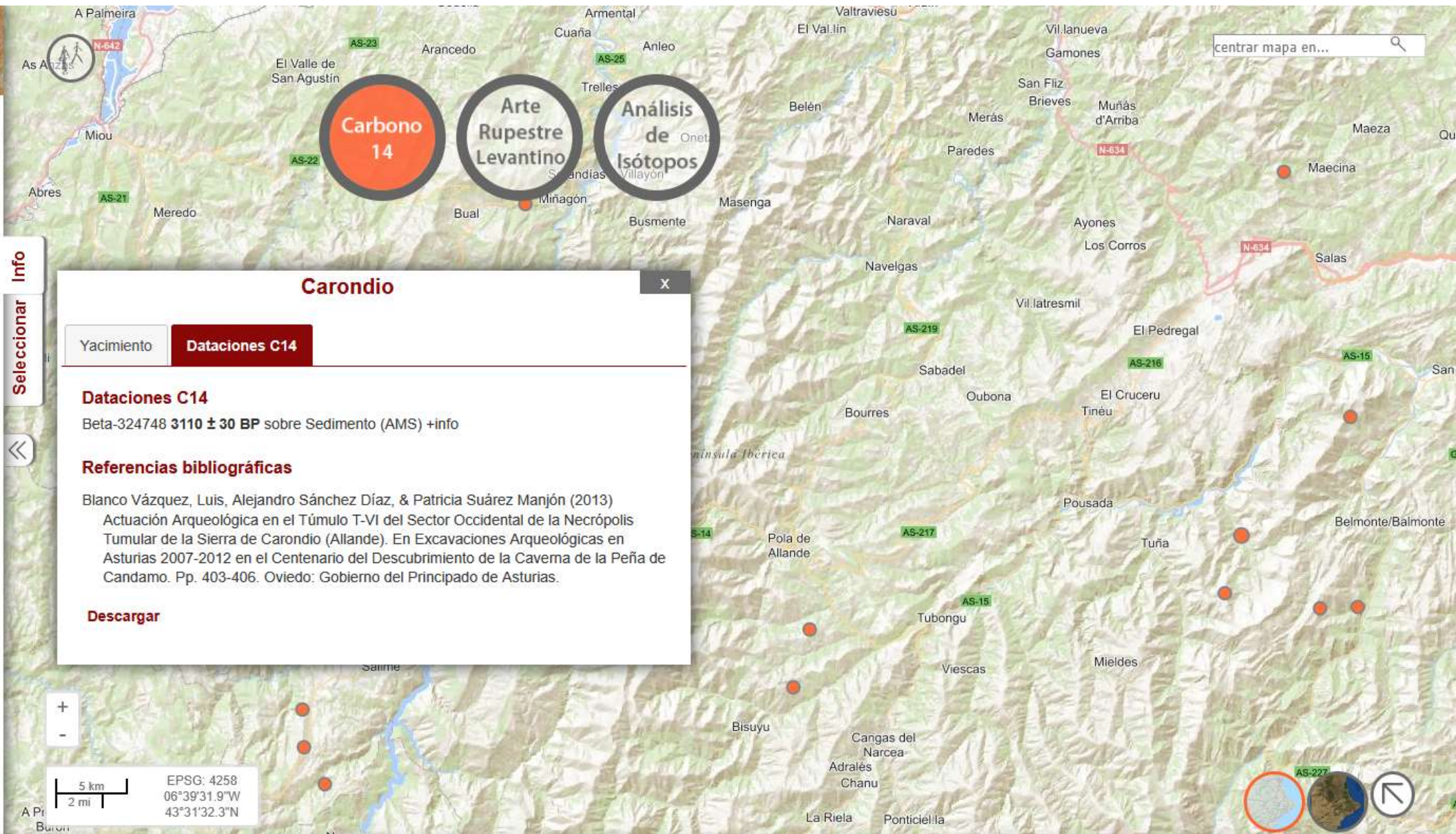
Blanco Vázquez, Luis, Alejandro Sánchez Díaz, & Patricia Suárez Manjón (2013)
Actuación Arqueológica en el Túmulo T-VI del Sector Occidental de la Necrópolis
Tumular de la Sierra de Carondio (Allande). En Excavaciones Arqueológicas en
Asturias 2007-2012 en el Centenario del Descubrimiento de la Caverna de la Peña de
Candamo. Pp. 403-406. Oviedo: Gobierno del Principado de Asturias.

Descargar



5 km
2 mi

EPSG: 4258
06°39'31.9"W
43°31'32.3"N



Retrieval-Augmented Generation (RAG)

Combina la capacidad generativa de los grandes modelos de lenguaje (LLM) con la precisión y actualización de bases de datos externas especializadas, creando un sistema de consulta científica.

Limitaciones de los *Large Language Models* (LLM)



Alucinaciones

Los LLM pueden generar respuestas que parecen coherentes pero realmente son incorrectas o inventadas, comprometiendo la fiabilidad de la información.



Conocimiento desactualizado

El conocimiento de los modelos deja de actualizarse tras su entrenamiento, sin acceso a información reciente o actualizaciones del dominio.



Dominios especializados

En campos como la arqueología, los LLM carecen del conocimiento profundo y específico necesario para respuestas precisas y contextualizadas.

Retrieval-Augmented Generation (RAG)



Acceso a bases de datos externas



Mayor fiabilidad



**Reducción de costes
computacionales**

Retrieval-Augmented Generation (RAG)



Acceso a bases de datos externas



Mayor fiabilidad



**Reducción de costes
computacionales**

Retrieval-Augmented Generation (RAG)



Acceso a bases de datos externas



Mayor fiabilidad

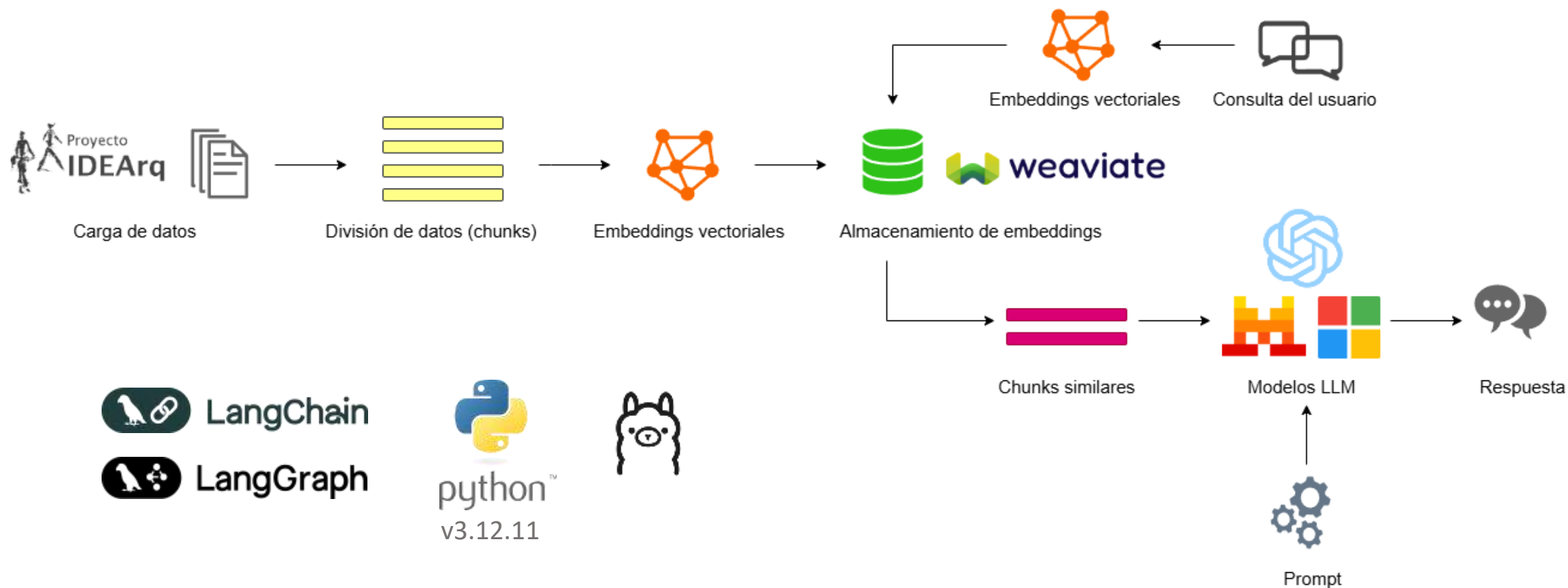


**Reducción de costes
computacionales**

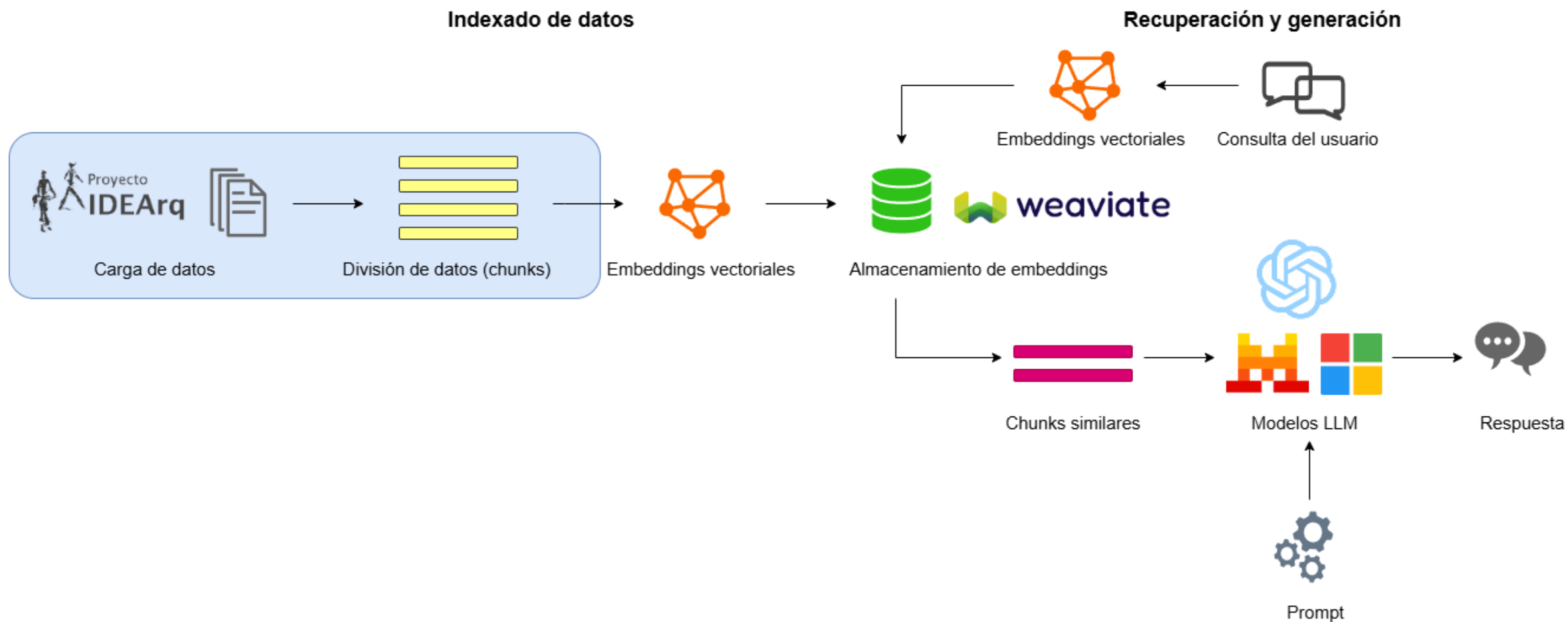
RAG – Arquitectura del sistema

Indexado de datos

Recuperación y generación



RAG – Indexación



RAG – Indexación → Carga de datos



➤ Corpus de publicaciones en formato **PDF** de IDEArq.

- 525 artículos (2000-2024).
- Español, catalán, inglés, portugués y francés.
- Artículos, tesis doctorales, etc.
- Temáticas variadas en arqueología.

➤ Capa de dataciones C^{14} en formato **GeoJSON**.

➤ Tabla en formato PDF de dataciones de C^{14} de la Península Ibérica.

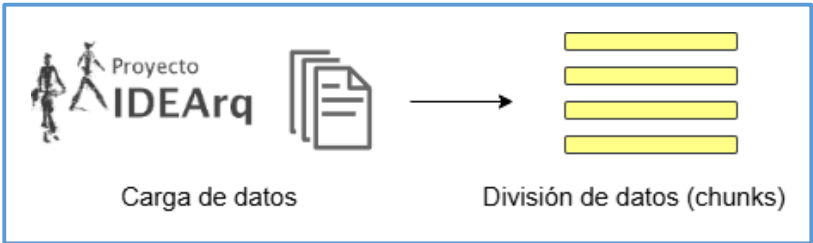


Tabla 1. Muestra de dataciones de C^{14}

Fecha	Referencia	Tipo de muestra	Material	Método	Yacimiento	Ubicación
30 BP	KIA-4010/29816	Referido	CarbónPinaceae - Pinus		Estany de la Coveta	ES)
4460 ± 120 BP	GifA-98175	No Referido	Madera / Carbón	AMS	Cueva de la Pileta	Benaoján (Málaga, ES)
4460 ± 680 BP	GaK-2551	No Referido	Madera / Carbón	Método convencional	Cueva de la Lloseta	Ribadesella (Asturias, ES)
4460 ± 70 BP	Beta-90621	No Referido	Madera / Carbón	Método no referido	Bauma del Serrat del Pont	Tortellà (Girona, ES)
4460 ± 40 BP	Sac-2963	Por Asignar	ConchaMytilidae - Mytilus	Método convencional	São Julião	Carvoeira (Lisboa, PT)
4454 ± 56 BP	AA-78191	No Referido	HuesoBovidae - Ovis/ Capra	AMS	Cueva de Arangas	Cabrales (Asturias, ES)

RAG – Indexación → División de datos (*chunks*)

¿Qué es?

- Fragmentación de los documentos en secciones pequeñas (*chunks*).
- No se almacenan los textos completos.

¿Por qué se hace?

- **Mejorar la recuperación:** el modelo accede solo a los fragmentos relevantes.
- **Evitar ruido** y optimizar recursos.

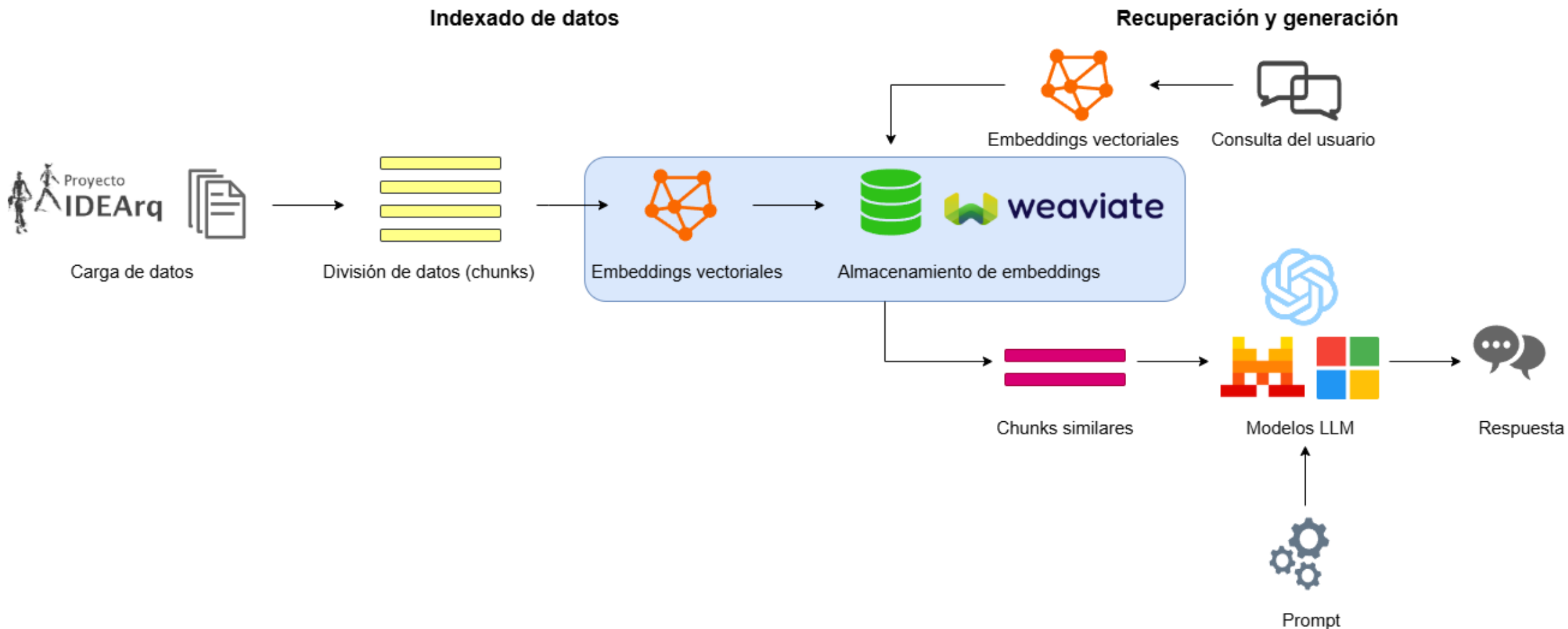
Equilibrio del tamaño

- *Chunks* grandes → más contexto, más ruido.
- *Chunks* pequeños → menos ruido, posible pérdida de información.

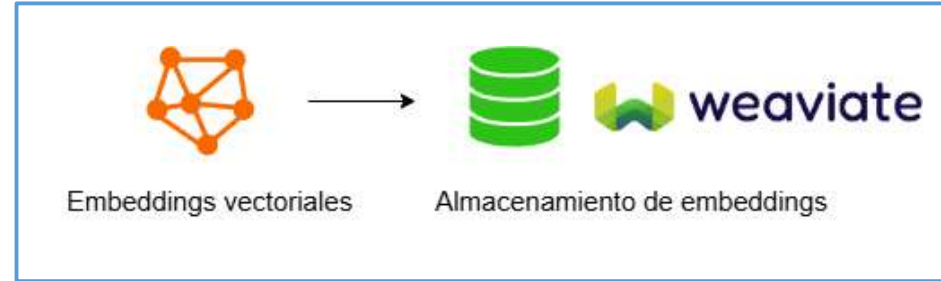
```
CHUNK_CONFIG = {  
    "chunk_size": 800,  
    "chunk_overlap": 50,  
    "length_function": len,  
    "add_start_index": True,  
    "separators": ["\n\n\n", "\n\n", "\n", ". ", ", ", " ", ""],  
}
```

Figura 1. Estrategia para dividir los documentos de texto en *chunks*

RAG – Indexación → *Embeddings* vectoriales



RAG – Indexación → *Embeddings* vectoriales



Massive Text Embedding Benchmark (MTEB) leaderboard



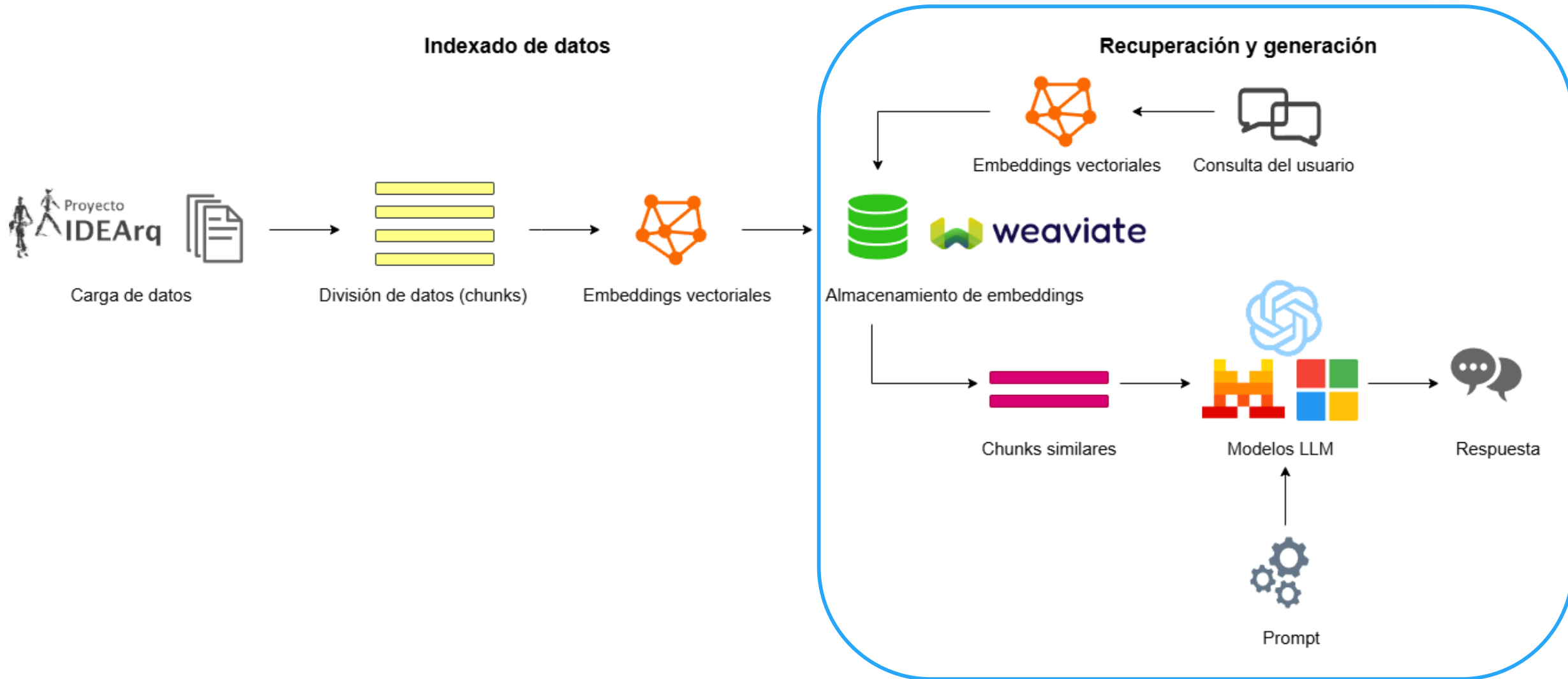
Qwen3-
Embedding

~8B

Linq-Embed-
Mistral

~7B

RAG – Recuperación y generación



RAG – Prompt

""Eres un asistente experto en arqueología, historia y sistemas de información geográfica. Responde la pregunta usando los contextos proporcionados, que pueden estar en español, inglés, francés, catalán o portugués.

Sintetiza información de todos los contextos relevantes independientemente de su idioma. Si encuentras información relevante en cualquier idioma, úsala para construir tu respuesta en el mismo idioma en el que se realiza la pregunta.

Instrucciones:

- Sintetiza información de todos los contextos relevantes, independientemente del idioma.
- Redacta la respuesta en el mismo idioma en que se formula la pregunta.
- Integra la información parcial si aporta valor, incluso si no es completa.
- Si no hay información relevante, responde literalmente: "No hay información suficiente en el contexto".
- No inventes datos no presentes en los contextos.
- Siempre que sea posible, cita datos concretos del contexto (autores, fechas, publicaciones, yacimientos, cronologías, dataciones, coordenadas, etc.).
- Mantén un tono académico, claro y estructurado, orientado a un público de investigación.

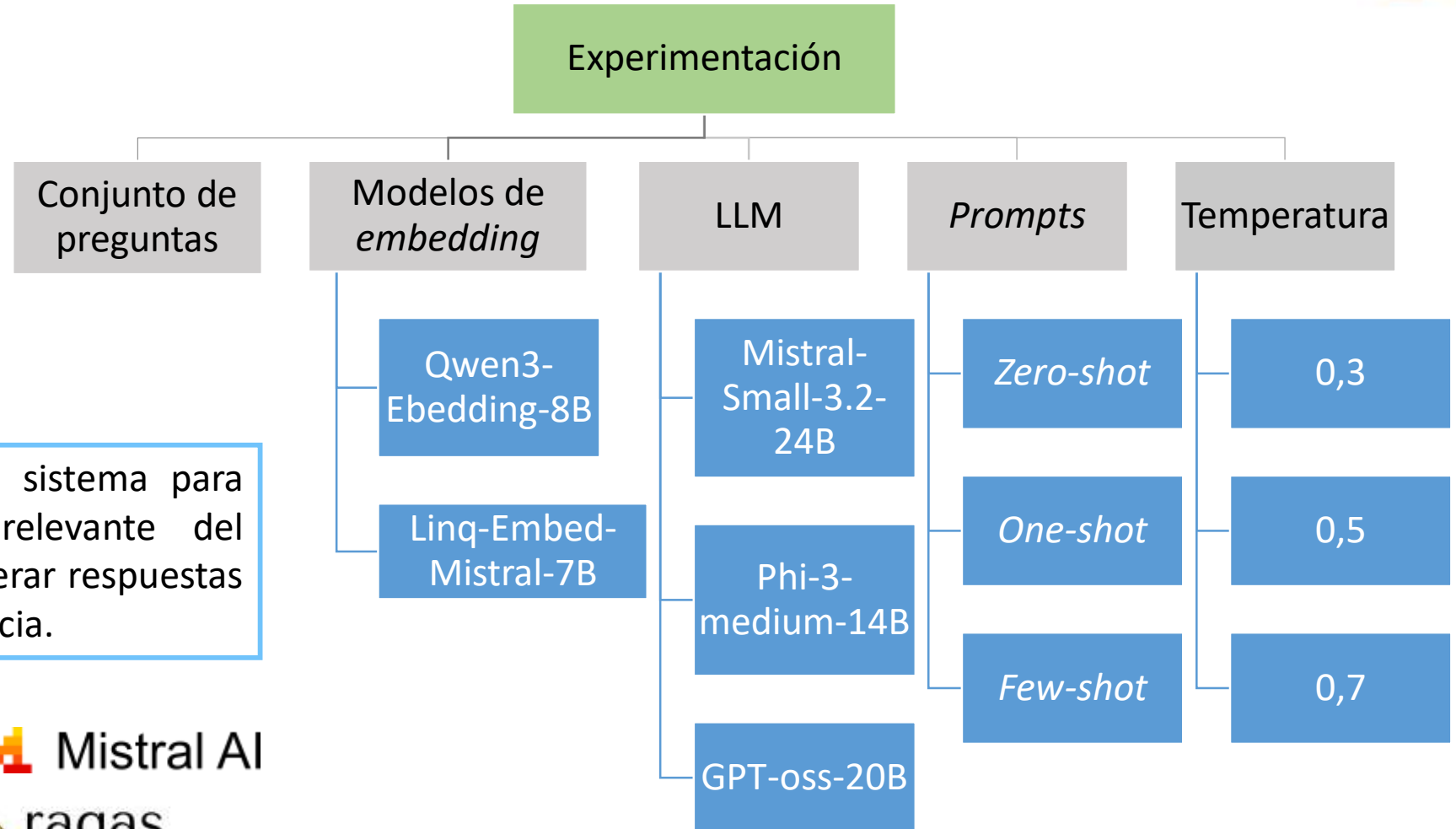
CONTEXTO: {context}

PREGUNTA: {question}

Basándote en la información anterior y en tu conocimiento sobre arqueología y sistemas de información geográfica, responde:

RESPUESTA: ""

Evaluación del sistema RAG



Evaluar la capacidad del sistema para recuperar información relevante del corpus documental y generar respuestas fundamentadas en evidencia.

Evaluación del sistema RAG - Métricas

Context precision

- Mide cómo de relevante es el contexto recuperado para responder a una pregunta.

Context recall

- Evalúa si el sistema encontró todos los datos relevantes presentes en la base de conocimientos (todo el contexto).

Faithfulness

- Cuantifica el grado de alucinación de la respuesta generada.



Answer correctness

- Mide la similitud semántica entre la respuesta generada por el modelo de lenguaje y la respuesta de referencia (*ground truth*), utilizando los modelos de *embedding*.

Rango [0, 1]

GPT-oss-20B

Pregunta: ¿Cuál es el yacimiento con la fecha de C14 más antigua de las Islas Baleares?

1. Sitio con la fecha más antigua:

- El registro más antiguo y claramente antropogénico en las Islas Baleares es ****4165±30B.P.**** obtenido de hueso en la ****Cova de Moleta / Moleta Petita**** (Sòller).
- La fecha proviene de la certificación KIK/KIA1833/14003 (categorizada como “sobre Hueso”) y se publica en la metodología de la ****Bada-VII**** como documento de referencia sobre la colonización del archipiélago.

2. Contexto de otras fechas relevantes:

- Se listan otras dataciones que, aunque relativamente antiguas, son posteriores a la Cova de Moleta (p. ej. 4135 B.P. en Cova de Moleta, 8000 B.P. en Cova de Canet).
- Datos de 8000 y 5800 B.P. reportados en la Cova de Moleta y en la Cova de Canet son considerados ****sin resultado**** o ****contextos naturales****, por lo que no se incluyen como datos fiables de presencia humana..

Pregunta: ¿Cuáles son las fechas más antiguas de la extracción de sílex en la Península Ibérica?

1. **São Pedro de Canaferrim (Lisboa, Portugal, 7750±50 B.P.).** Esta es la fecha más antigua registrada para la extracción de sílex en la península ibérica, según el registro de “Beta164712 7750±50 B.P.” en “Madera/Carbón” del sitio.
2. **Cantarranas (Cádiz, 5490±50 B.P.).** Las dataciones de “Sac1659 5490±50 B.P.” sobre concha en el taller lítico son la segunda más antigua.
3. **La Profunda (León, 4075±35 B.P.).** Las muestras “Ua35780 4075±35 B.P.” y “Ua35779 3950±35 B.P.” provienen de esta mina de extracción de sílex.

Pregunta: Principales yacimientos de la Segunda Edad del Hierro en la provincia de León.

Peña del Castro

Pregunta: Yacimientos calcolíticos de la Península Ibérica en los que se han hallado objetos de marfil.

No hay información suficiente en el contexto proporcionado para identificar yacimientos calcolíticos de la Península Ibérica donde se hayan hallado objetos de marfil.

Producción y consumo de marfil en la Edad del Bronce: Cabezo Redondo (Villena, Alicante) como marco

Production and consumption of ivory during the Bronze Age: Cabezo Redondo (Villena, Alicante) as a framework

VIRGINIA BARCIELA GONZÁLEZ
I.U.I. en Arqueología y Patrimonio Histórico
Universidad de Alicante
Carretera San Vicente, s/n. 03690 San Vicente del Raspeig (Alicante)
virginia.barcia@ua.es
<https://orcid.org/0000-0002-9623-8982>

JUAN ANTONIO LÓPEZ PADILLA
Museo Arqueológico de Alicante –MARQ–
Plaza Doctor Gómez Ulla, s/n. 03013 Alicante
japadi@diputacionalicante.es
<https://orcid.org/0000-0002-1506-4731>

GABRIEL GARCÍA ATIÉNZAR
I.U.I. en Arqueología y Patrimonio Histórico
Universidad de Alicante
Carretera San Vicente, s/n. 03690 San Vicente del Raspeig (Alicante)
g.garcia@ua.es
<https://orcid.org/0000-0001-9390-8111>

MAURO S. HERNÁNDEZ PÉREZ
Universidad de Alicante
mauro.hernandez@ua.es
<https://orcid.org/0000-0001-8225-4730>

Pregunta: ¿Cuál es el yacimiento calcolítico de la provincia de Jaén más alejado de la ciudad de Jaén?

Para determinar el yacimiento más lejano a la ciudad de Jaén, primero debemos conocer las coordenadas geográficas de Jaén. La ciudad de Jaén se encuentra aproximadamente en las coordenadas (37.7673, -3.7878).

El contexto proporcionado incluye cuatro puntos con sus respectivas coordenadas:

1. POINT (-3.78521246922312, 37.9359566644291)
2. POINT (-3.546180256079071, 37.928296085760316)
3. POINT (-3.60226346991079, 37.828879825305556)
4. POINT (-3.364525258986248, 38.010115119339105)

Para transformar estas distancias de grados a kilómetros, recordemos que 1 grado de latitud es aproximadamente igual a 111 km y 1 grado de longitud en el ecuador también es aproximadamente 111 km. Sin embargo, esta conversión varía ligeramente con la latitud. Dado que Jaén se encuentra en una latitud moderada (37.7673°N), podemos usar un factor de conversión cercano al promedio para latitudes medias.

Por lo tanto, el yacimiento más lejano a la ciudad de Jaén es el punto 4, con una distancia aproximada de 54.17 kilómetros, que corresponde a Eras del Alcázar.

¿Cuál es el yacimiento calcolítico de la provincia de Jaén que está más alejado de la ciudad de Jaén?



Pensó durante 3m 35s >

Buenísimo — voy al grano y con datos. He buscado los yacimientos calcolíticos de la provincia de Jaén con coordenadas públicas y he calculado la distancia en línea recta (haversine) desde el **centro de la ciudad de Jaén**. Resultado (mejor tratarlo como *orientativo*, ver *aclaraciones* al final):

Resumen corto (respuesta)

Con los datos públicos disponibles, **los yacimientos situados en la parte noreste de la provincia (Sierra de Cazorla / términos de Hinojares — Pozo Alcón)** son los que quedan más alejados de la ciudad de Jaén.

Ejemplos representativos:

- **Pozo Alcón / zona del valle del Guadiana Menor (yacimientos recientes/dolmenes)** — ~75.6 km en línea recta desde Jaén. [Vive Más Andal... +1](#)
- **Castellones de Ceal (Hinojares, Sierra de Cazorla)** — ~65.4 km en línea recta (nota: Castellones de Ceal es un asentamiento con fases ibéricas/argárico; no es estrictamente “calcolítico” en su cronología principal, pero aparece en los inventarios arqueológicos de la zona). [Castillos y Fortal... +1](#)

Para comparar, algunos yacimientos calcolíticos bien documentados más próximos a la ciudad:

- **Venta del Rapa (Mancha Real)** — ~17.3 km. [Wikipedia](#)
- **Cueva / yacimientos en Bedmar (Sierra Mágina)** — ~33.8 km. [Wikipedia +1](#)
- **Cerro de los Vientos (Puente del Obispo, recinto de fosos calcolítico)** — ~30.1 km. [Webgea +1](#)

Conclusiones

1. Esta integración entre IDE e Inteligencia Artificial permite establecer conexiones entre datos espaciales, textos científicos y conceptos arqueológicos, abriendo nuevas vías para la investigación interdisciplinar y la exploración del patrimonio.
2. El sistema permite explotar los recursos de conocimiento arqueológico de IDEArq, facilitando el acceso a información especializada a través de consultas en lenguaje natural.
3. Diseñar un chatbot.
4. Métodos para mejorar el rendimiento:
 - Optimizar el indexado mediante el desarrollo de un grafo de conocimiento como índice para aplicar una estructura más jerárquica que establezca conexiones entre conceptos y entidades, reduciendo las alucinaciones mediante relaciones semánticas explícitas.
 - Implementar técnicas de *re-ranking* para optimizar la posición de *chunks* relevantes en el contexto.

Elena Aguayo Jara
elena.aguayo@cchs.csic.es

■ **Muchas
gracias**

■ **Muito
obrigado**

XVI Jornadas Ibéricas de Infraestructura de Datos Espaciales

Resumen y video disponible en www.idee.es/web/idee/jiide

IDEArq – Corpus de Pintura Rupestre Levantina (CPRL)

- Fotografías realizadas entre 1969 y 1974.
- Periodo **Postpaleolítico**.
- Más de **3000** imágenes.

Los Grajos / Barranco de los Grajos (Abrigo Grande)

Cieza (Murcia, ES)
Descriptores:
Hábitat, Hábitat rupestre (Neolítico)
Arte rupestre, Abrigo (Sin atribución cronológica)

Los Grajos / Barranco de los Grajos (Abrigo Grande)

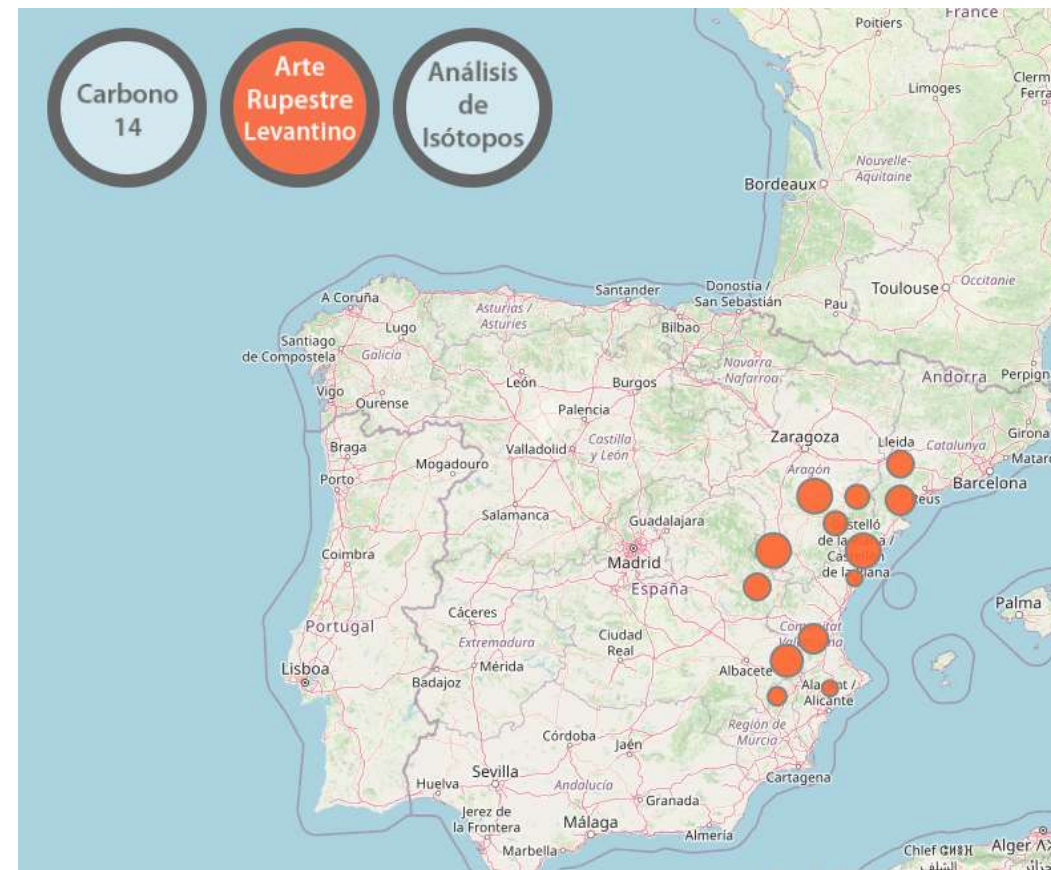
Yacimiento Dataciones C14 **Corpus de Pintura Rupestre Levantina**

Corpus de Pintura Rupestre Levantina

Fotos de conjunto

Abrigo I (cavidad dcha.). Imágenes disponibles (octubre 1975):

- Figura 2, Antropomorfo
- Figura 4, Cuadrúpedo Indeterminado
- Figura 5, Antropomorfo
- Figura 7, Indeterminado; Figura 8, Antropomorfo; Figura 9, Antropomorfo; Figura 10, Antropomorfo; Figura 6, Antropomorfo; Figura 11, Antropomorfo
- Figura 6, Antropomorfo
- Figura 8, Antropomorfo
- Figura 8, Antropomorfo
- Figura 8, Antropomorfo
- Figura 9, Antropomorfo; Figura 11, Antropomorfo; Figura 10, Antropomorfo
- Figura 9, Antropomorfo
- Figura 9, Antropomorfo
- Figura 9, Antropomorfo
- Figura 9, Antropomorfo
- Figura 10, Antropomorfo



IDEArq – Corpus de Pintura Rupestre Levantina (CPRL)

- Fotografías realizadas entre 1969 y 1974.
- Periodo **Postpaleolítico**.
- Más de **3000** imágenes.

Los Grajos / Barranco de los Grajos (Abrigo Grande)

Cieza (Murcia, ES)

Descriptores:

Hábitat, Hábitat rupestre (Neolítico)

Arte rupestre, Abrigo (Sin atribución cronológica)

Info

Seleccionar

<<

RM-B35

B36

Los Grajos / Barranco de los Grajos (Abrigo Grande)

Yacimiento

Dataciones C14

Corpus de Pintura Rupestre Levantina

Corpus de Pintura Rupestre Levantina

Fotos de conjunto

Abrigo I (cavidad dcha.). Imágenes disponibles (octubre 1975):

Figura 2, Antropomorfo

Figura 4, Cuadrúpedo Indeterminado

Figura 5, Antropomorfo

Figura 7, Indeterminado; Figura 8, Antropomorfo; Figura 9, Antropomorfo; Figura 10, Antropomorfo; Figura 6, Antropomorfo; Figura 11, Antropomorfo

Figura 6, Antropomorfo

Figura 6, Antropomorfo

Figura 8, Antropomorfo

Figura 8, Antropomorfo

Figura 8, Antropomorfo

Figura 9, Antropomorfo; Figura 11, Antropomorfo; Figura 10, Antropomorfo

Figura 9, Antropomorfo

Figura 9, Antropomorfo

Figura 9, Antropomorfo

Figura 9, Antropomorfo

Figura 10, Antropomorfo



IDEArq – Archivo de valores isotópicos de la Península Ibérica (δimP)

- Valores isotópicos de **estroncio y oxígeno** obtenidos de restos óseos humanos y animales.
- Reconstrucción de la dieta y los patrones de movilidad de las poblaciones prehistóricas.
- Datos de **143 yacimientos** y **1938 análisis sobre apatita**.
- Análisis de colágeno.

