La cocina de los grandes modelos del lenguaje aplicada al procesamiento de información sobre biodiversidad: exprimiendo GBIF

José Bengoa* y Sofia Turkina**

** Junta de Castilla y León ** Universidad Complutense de Madrid*

Jornadas sobre Información de Biodiversidad y Administraciones Ambientales 2025

Resumen

Los grandes modelos de lenguaje (Large Language Models, LLM) han revolucionado el procesamiento del lenguaje natural mediante técnicas de vectorización semántica que capturan relaciones complejas entre palabras y conceptos. Este trabajo propone la aplicación de estas mismas técnicas al análisis de datos de biodiversidad, estableciendo una analogía fundamental: donde el procesamiento del lenguaje natural utiliza palabras y frases, el procesamiento de biodiversidad puede utilizar especies y comunidades. Presentamos una metodología denominada Biodiversity Vectorization (BV) que transforma datos de ocurrencias de GBIF en representaciones vectoriales (embeddings), permitiendo operaciones matemáticas sobre entidades biológicas que preservan relaciones ecológicas y biogeográficas. Se comentan algunas aplicaciones más inmediatas como detección de anomalías, predicción de hábitats, identificación de áreas poco muestreadas y generación de mapas de similitud ecológica. Este enfoque abre nuevas posibilidades para el análisis de grandes volúmenes de datos de biodiversidad, aprovechando décadas de avances en procesamiento del lenguaje natural y aprendizaje automático.

Palabras clave: Modelos de lenguaje, Biodiversity Vectorization, GBIF, Word2Vec, embeddings, análisis ecológico, Natural Language Processing, biodiversidad

Abstract

Large Language Models (LLMs) have revolutionized natural language processing through semantic vectorization techniques that capture complex relationships between words and concepts. This work proposes the application of these same techniques to biodiversity data analysis, establishing a fundamental analogy: where natural language processing uses words and sentences, biodiversity processing can use species and communities. We present a methodology called Biodiversity Vectorization (BV) that transforms GBIF occurrence data into vector representations (embeddings), enabling mathematical operations on biological entities that preserve ecological and biogeographical relationships. Results demonstrate practical applications in anomaly detection, habitat prediction, identification of under-sampled areas, and generation of ecological similarity maps. This approach opens new possibilities for analyzing large volumes of biodiversity data, leveraging decades of advances in natural language processing and machine learning.

Keywords: Language models, Biodiversity Vectorization, GBIF, Word2Vec, embeddings, ecological analysis, Natural Language Processing, biodiversity

1. Introducción

Estas jornadas constituyen un foro privilegiado de intercambio entre las administraciones gestoras de la información de biodiversidad y el mundo académico. En el espíritu de este encuentro, el presente trabajo nace de una colaboración entre dos perspectivas complementarias: la gestión práctica de datos de biodiversidad y el rigor matemático de la academia, representadas en la colaboración con Sofía Turkina, estudiante de matemáticas de la Universidad Complutense de Madrid.

Somos testigos y protagonistas de una revolución tecnológica sin precedentes. Existe un antes y un después del 30 de noviembre de 2022, fecha en que OpenAI lanzó ChatGPT al público general. De hecho, en las jornadas que celebramos en Vitoria en 2022, una semana antes de ese momento histórico, estuvimos hablando de GPT3, de las impresionantes capacidades de algunos modelos generativos y de las potenciales aplicaciones en el ámbito de la biodiversidad.

La evolución de la percepción social sobre la inteligencia artificial en estos tres años puede sintetizarse en tres fases: la primera, el momento de asombro inicial (*"wow moment"*); la segunda, la fase de normalización y crítica (*"eso ya no es una novedad, además no hace bien esta o aquella tarea"*); y la tercera, la fase de aplicación práctica (*"¿qué puede hacer el LLM por mí?"*). En la esfera específica de los LLM, hemos sido testigos de importantes hitos: la multimodalidad (2023), las capacidades de razonamiento avanzado (2024) y, actualmente, el desarrollo de agentes autónomos (2025).

El presente trabajo se centra en una aplicación concreta: la transferencia de técnicas de procesamiento del lenguaje natural al análisis de datos de biodiversidad. Para comprender el alcance de esta propuesta, es necesario primero establecer el marco conceptual y técnico en el que se desarrolla.

2. La Revolución de los Modelos de Lenguaje: Contexto y Capacidades

2.1. Capacidades Conversacionales vs. Capacidades Agénticas

Es fundamental distinguir entre un modelo que *solo habla* y uno que *hace cosas*. Cuando un modelo puede ejecutar acciones, decimos que tiene capacidades agénticas. A finales de 2024, Anthropic publicó el Model Context Protocol (MCP), un estándar que ha sido adoptado por la industria, incluyendo OpenAI, para facilitar la integración de estas capacidades.

Este desarrollo abre dos dimensiones fundamentales. En primer lugar, permite que las inteligencias artificiales trabajen de forma autónoma en tareas delegadas. En segundo lugar, y quizás más profundo, reconoce que el lenguaje no es solo un depositario de conocimiento, sino también nuestra herramienta fundamental para comunicar y razonar. ¿Qué son las ideas que no se expresan en palabras? ¿meras secuencias de sinapsis?

2.2. El Lenguaje Natural como Interfaz Universal

La consecuencia práctica más inmediata es que la relación con las aplicaciones informáticas y, en general, con las máquinas, está incorporando el lenguaje natural como interfaz universal. Ya no será necesario aprender lenguajes de programación específicos o interfaces complejas; bastará con expresar nuestras intenciones en lenguaje natural a una IA con capacidades agénticas.

Esta transformación ya es una realidad consolidada en el mundo del desarrollo de software. Prácticamente ningún desarrollador profesional trabaja sin asistencia de IA. El paradigma tradicional de *"picar código"* está siendo sustituido por la programación mediante instrucciones en lenguaje natural. Pero esta revolución se extenderá a todos los ámbitos. En el campo de la cartografía, por ejemplo, ya hemos desarrollado complementos de QGIS que atienden a instrucciones usando el protocolo MCP. Esta iniciativa está en fase experimental, pero no por razones técnicas, ya que las herramientas están disponibles y su implementación no es compleja, sino por razones más bien operativas.

2.3. Permisos y Control: La Pendiente Inevitable

Las capacidades agénticas requieren que otorguemos permisos a los LLM para acceder a nuestros recursos (archivos, correo electrónico, bases de datos) y para ejecutar acciones (ejecutar código, modificar documentos, realizar consultas). Esto plantea evidentes cuestiones de seguridad y privacidad que deben estar rigurosamente controladas. No obstante, es una pendiente por la que la sociedad avanzará inexorablemente, por mucho que inicialmente nos resistamos.

3. Fundamentos Técnicos de los Modelos de Lenguaje

3.1. Fuentes de Conocimiento de un LLM

Para comprender las posibilidades y limitaciones de los LLM, es necesario entender de qué fuentes extraen su capacidad de respuesta. Un modelo de lenguaje se nutre de seis componentes principales:

- **El modelo preentrenado:** Un archivo de gran tamaño (desde 5-15 GB para modelos pequeños hasta más de 350 GB en casos como GPT-3 o superiores) que contiene los pesos neuronales resultantes del entrenamiento con enormes corpus de texto. Este preentrenamiento proporciona capacidades genéricas de comprensión y generación de lenguaje.
- **El postentrenamiento (RLHF):** Mediante Aprendizaje por Refuerzo con Retroalimentación Humana (Reinforcement Learning with Human Feedback), se ajusta el modelo para que adopte comportamientos específicos, desarrolle una *"personalidad"* coherente y respete normas de conducta.
- **El prompt del sistema:** Instrucciones ocultas al usuario que definen el comportamiento del modelo. Por ejemplo: *"Eres un modelo de lenguaje entrenado por OpenAI. Responde de manera concisa y clara. Sé cortés en todo momento"*.
- **Capas de seguridad (guardrails):** Mecanismos que filtran tanto las peticiones entrantes como las respuestas generadas, asegurando que el modelo no se desvíe de su función ni viole restricciones éticas o legales.
- **El prompt del usuario:** La consulta o instrucción específica que el usuario proporciona.

- **El contexto:** Aquí reside *"la madre del cordero"*. El contexto incluye toda la información que el modelo puede considerar para generar su respuesta: conversación previa, documentos adjuntos, datos recuperados, etc.

3.2. La Ventana de Contexto: Evolución y Limitaciones

Los modelos Transformer revolucionaron el procesamiento del lenguaje natural precisamente porque superaron las limitaciones de memoria de arquitecturas anteriores. Los mecanismos de atención permitieron escapar del estancamiento en que se encontraba el campo (Vaswani et al, 2017).

La evolución de la ventana de contexto ha sido espectacular: de 4.000 tokens iniciales se ha pasado a 32.000 (Mistral Large), 65.000 (Mistral 8x22B), 128.000 (GPT-4o), 200.000 (Claude 3.5) y hasta 1.000.000 de tokens (Gemini 1.5). Usando la equivalencia aproximada de 1 página = 400 tokens, o 200.000 tokens = 500 páginas, estamos hablando de contextos que pueden abarcar hasta 2.500 páginas.

Aunque estas cifras parecen impresionantes, la ventana de contexto sigue siendo uno de los principales desafíos actuales. Si el contexto pudiera ser infinito, las capacidades de los LLM se multiplicarían exponencialmente.

3.3. RAG: Trabajando con Grandes Volúmenes de Información

Para superar las limitaciones de la ventana de contexto, se ha desarrollado la técnica de Generación Aumentada por Recuperación (Retrieval-Augmented Generation, RAG). Este enfoque no intenta meter toda la información en el contexto, sino que busca y recupera selectivamente los fragmentos más relevantes para cada consulta específica, insertándolos dinámicamente en el contexto del modelo.

RAG se basa fundamentalmente en la vectorización semántica: los documentos se dividen en fragmentos, cada fragmento se transforma en un vector mediante un modelo de embedding, y posteriormente se realiza una búsqueda de similitud entre el vector de la consulta del usuario y los vectores de los fragmentos almacenados. Los fragmentos más similares se recuperan y se añaden al contexto antes de generar la respuesta.

4. Del Procesamiento del Lenguaje Natural a la Biodiversidad

4.1. Las Palabras como Vectores

En el procesamiento del lenguaje natural, cada palabra se representa como un vector de alta dimensión en un espacio semántico. Las frases son ecosistemas de palabras que componen paisajes no tanto visuales como conceptuales o dialécticos. Esta representación vectorial no es arbitraria: se aprende mediante modelos entrenados con grandes corpus de texto, de manera que las palabras con significados o usos similares quedan próximas en el espacio vectorial.

Las impresionantes capacidades de los LLM derivan precisamente de esta vectorización (*embedding*), realizada por modelos especializados que asignan vectores en función de relaciones de afinidad, vecindad y relación semántica. Así, *"casa"* y *"puerta"* estarán más cerca entre sí que *"rueda"* y *"árbol"*.

Para capturar estas relaciones complejas se requieren espacios de alta dimensionalidad: no 3, 12 o 24 dimensiones, sino centenares o miles. Ejemplos de modelos de embedding incluyen all-MiniLM-L6-v2 (384 dimensiones), BERT (768 dimensiones), Cohere (1.024 dimensiones) y text-embedding-3-large de OpenAI (3.072 dimensiones).

4.2. La Analogía Fundamental: Especies como Tokens

En este punto, aparentemente alejados del tema inicial de GBIF y la biodiversidad, se revela la conexión fundamental. Basta con realizar una traducción conceptual:

- Donde decíamos ***"palabras"***, digamos ***"especies"***
- Donde decíamos ***"frases"*** o ***"documentos"***, digamos ***"comunidades"*** y ***"ecosistemas"***
- Donde decíamos ***"contexto semántico"***, digamos ***"contexto ecológico y biogeográfico"***

Podemos tratar a las especies como vectores, exactamente igual que tratamos a las palabras. Y podemos hacer lo mismo con comunidades y ecosistemas.

4.3. Del NLP al VBP: Vectorial Biodiversity Processing

La ecología y las matemáticas han colaborado históricamente, pero hasta ahora estos fueron juegos de corto alcance, con capacidad limitada para modelizar relaciones complejas: Análisis de Componentes Principales (PCA), TWINSPAN, MaxEnt, etc. Los LLM nos han enseñado que se puede trabajar con entidades como palabras o tokens mediante vectores obtenidos con entrenamiento que captura la semántica y las relaciones entre ellos, y posteriormente pedirles que revelen lo que han aprendido.

Proponemos aplicar las mismas técnicas del Procesamiento del Lenguaje Natural (Natural Language Processing, NLP) al Procesamiento Vectorial de la Biodiversidad (Vectorial Biodiversity Processing, VBP). Tratar a las especies como códigos meramente cualitativos en ecología no aprovecha todo el potencial de las herramientas matemáticas a nuestra disposición. La incorporación de la semántica en la vectorización es clave para esta tarea.

5. Metodología: Biodiversity Vectorization

5.1. Fundamentos Conceptuales

La pregunta natural es: ¿cómo se implementa esto en la práctica? La respuesta sigue el mismo patrón que en NLP:

- **Necesitamos un corpus:** Un conjunto amplio de datos de biodiversidad estructurados de manera apropiada.
- **Necesitamos un modelo:** Que entrenaremos para asignar vectores a cada entidad biológica (especies, localidades, comunidades, ecosistemas).
- **Necesitamos definir la unidad de análisis:** ¿Vectorizamos especies? ¿Citas individuales? ¿Inventarios? ¿Localizaciones? ¿Ecosistemas? Cada nivel ofrece posibilidades diferentes.

5.2. GBIF como Corpus Fundamental

Para no dejar la propuesta en el plano meramente teórico y cerrar el círculo volviendo a la referencia inicial, GBIF (Global Biodiversity Information Facility) proporciona un corpus formidable. Con más de 2.000 millones de registros de ocurrencias de especies de todo el mundo, estandarizados y accesibles mediante una API robusta, GBIF constituye exactamente el tipo de conjunto de datos que necesitamos.

La construcción del corpus sigue estos pasos:

- **Descarga de ocurrencias:** Mediante la API de GBIF (utilizando pygbif (Chamberlain et al., 2024) o llamadas HTTP directas), descargamos registros filtrados por criterios de calidad (coordenadas válidas, sin flags críticos).
- **Agregación espacial:** Los datos se agrupan en celdas de una rejilla espacial regular (por ejemplo, cuadrículas de 0.1° x 0.1°).
- **Construcción de frases ecológicas:** Cada celda se convierte en una *"frase"* y las especies presentes en ella son las *"palabras"*.

Por ejemplo:

> *Celda A → [“Narcissus pseudonarcissus”, “Trifolium pratense”, “Ranunculus acris”]*
>
> *Celda B → ["Pinus sylvestris", "Juniperus communis"]*

5.3. Entrenamiento del Modelo Word2Vec

Adoptamos este modelo a modo de ejemplo, por su sencillez para implementar los algunos ejemplos. Utilizamos la biblioteca Gensim (Rehurek & Sojka, 2010) para entrenar un modelo Word2Vec (Mikolov et al., 2013) sobre nuestro corpus ecológico. El modelo aprende:

- Semejanza ecológica entre especies basada en coocurrencia
- Patrones de coocurrencia espacial
- Afinidad ambiental entre celdas
- Gradientes biogeográficos implícitos en los datos

Los parámetros del modelo incluyen:

- **VECTOR_SIZE:** Dimensión de los vectores (típicamente 50-300)
- **WINDOW:** Tamaño de la ventana de contexto (especies vecinas a considerar)
- **MIN_COUNT:** Frecuencia mínima de aparición de una especie
- **EPOCHS:** Número de iteraciones de entrenamiento

5.4. Generación de Embeddings

Una vez entrenado el modelo, generamos dos tipos de embeddings:

- **Embedding de especies:** Directamente del modelo Word2Vec (Mikolov et al., 2013), cada especie obtiene un vector que representa su *"nicho semántico"* basado en las especies con las que coocurre.
- **Embedding de celdas:** Calculado como la media de los vectores de todas las especies presentes en cada celda. Este vector representa el *"perfil ecológico"* de la localidad.

5.5. Métricas de Similitud

Para comparar especies o celdas utilizamos la Similitud del Coseno, métrica estándar en análisis vectorial:

$$Similitud(A,B) = (A \cdot B) / (||A|| \times ||B||)$$

Donde A · B es el producto escalar y ||A||, ||B|| son las magnitudes (norma euclidiana) de los vectores. El resultado varía entre -1 (totalmente opuestos) y 1 (idénticos), con 0 indicando ortogonalidad (sin relación).

En el contexto ecológico:

- **Similitud cercana a 1:** Las celdas comparten composiciones de especies similares, sugiriendo condiciones ambientales o biogeográficas parecidas.
- **Similitud cercana a 0:** Las celdas tienen comunidades biológicas muy diferentes.
- **Similitud negativa:** Raramente ocurre, pero indicaría composiciones completamente opuestas.

6. Implementación Práctica con GBIF

6.1. Estructura del Script

Se ha desarrollado un script Python que implementa el flujo de trabajo descrito. El notebook está estructurado en las siguientes secciones:

- **Instalación de dependencias:** pygbif, gensim, scikit-learn, pandas, numpy, matplotlib, requests
- **Configuración de parámetros:** Reino, país, área geográfica, tamaño de celda, parámetros del modelo
- **Descarga de datos de GBIF:** Funciones para resolver nombres de especies y descargar ocurrencias
- **Construcción del corpus:** Agregación espacial y generación de
- *"frases ecológicas"*
- **Entrenamiento del modelo:** Word2Vec con Gensim (Rehurek & Sojka, 2010)
- **Generación de embeddings:** Para especies y celdas
- **Visualización:** PCA para reducción de dimensionalidad y mapas de similitud
- **Aplicaciones prácticas:** Casos de uso específicos detallados en la siguiente sección

6.2. Reproducibilidad y Acceso

El script es muy sencillo y únicamente pretende mostrar las posibilidades de la metodología propuesta y poner en evidencia la sencillez de las herramientas necesarias. Puede ejecutarse en local o en Google Colab, que es lo que usamos en las Jornadas para no tener que configurar un entorno. Los usuarios pueden:

- Modificar parámetros de búsqueda (reino, país, área geográfica)
- Ajustar la resolución espacial (tamaño de celda)
- Experimentar con diferentes arquitecturas de modelo
- Aplicar el análisis a taxones específicos

Esta disponible en github (https://github.com/jlbmdm/BiodiversityVectorProcessing*).*

7. Aplicaciones y Casos de Uso

La vectorización de biodiversidad permite múltiples aplicaciones prácticas de interés tanto para la investigación como para la gestión ambiental:

7.1. Detección de Citas Anómalas

Calculando la similitud entre el vector de una especie y el vector de las celdas donde ha sido citada, podemos identificar citas que no encajan con el perfil ecológico esperado. Una similitud muy baja entre una especie y la celda donde aparece puede indicar:

- Errores de identificación taxonómica
- Errores en las coordenadas geográficas
- Poblaciones relictas o introducidas de gran interés científico
- Cambios en la distribución debido a cambio climático

7.2. Predicción de Hábitats Esperables

Para una especie determinada, podemos calcular su vector de hábitat como la media de los embeddings de todas las celdas donde aparece. Posteriormente, calculamos la similitud de este vector con todas las celdas del territorio:

- Las celdas con alta similitud donde la especie no ha sido registrada son hábitats potencialmente favorables
- Esto puede orientar trabajos de prospección en áreas específicas
- También puede identificar áreas favorables para reintroducciones

7.3. Mapas de Similitud Ecológica

Generando mapas de calor basados en la similitud de embeddings, podemos visualizar:

- Patrones biogeográficos emergentes no evidentes en análisis tradicionales
- Gradientes ambientales implícitos en la distribución de especies
- Clústeres de celdas con composición similar que pueden corresponder a unidades biogeográficas o tipos de hábitat
- Transiciones y ecotonos entre diferentes regiones ecológicas

7.4. Análisis de Conectividad Ecológica

La similitud entre celdas puede interpretarse como una medida de conectividad ecológica, complementaria a métricas basadas en conectividad física:

- Identificación de corredores ecológicos basados en similitud de comunidades
- Detección de barreras ecológicas (zonas de baja similitud entre áreas adyacentes)
- Evaluación de fragmentación de hábitats desde una perspectiva funcional

8. Discusión

8.1. Ventajas del Enfoque

La aplicación de técnicas de NLP a datos de biodiversidad ofrece varias ventajas significativas:

- **Escalabilidad:** Los algoritmos están optimizados para grandes volúmenes de datos, precisamente lo que caracteriza a GBIF.
- **Captura de relaciones complejas:** Los embeddings pueden codificar patrones multidimensionales que serían difíciles de capturar con métodos tradicionales.
- **Aprendizaje no supervisado:** No requiere etiquetar los datos más allá de la información ya recogida en GBIF; el modelo aprende directamente de los patrones de coocurrencia.
- **Transferibilidad:** Las mismas técnicas pueden aplicarse a diferentes escalas espaciales, grupos taxonómicos o regiones geográficas.
- **Interpretabilidad:** Los vectores de similitud tienen interpretaciones ecológicas claras.

8.2. Limitaciones y Consideraciones

No obstante, es importante reconocer las limitaciones del enfoque:

- **Dependencia de la calidad de datos:** Los embeddings reflejarán sesgos presentes en GBIF (sesgos geográficos, taxonómicos, temporales).
- **Interpretación ecológica:** La coocurrencia no implica necesariamente interacción ecológica; puede reflejar simplemente tolerancias ambientales similares.
- **Resolución espacial:** El tamaño de celda elegido afecta significativamente a los resultados; debe justificarse ecológicamente.
- **Especies raras:** Especies con pocas citas pueden no generar embeddings robustos.
- **Variación temporal:** El método actual no incorpora explícitamente la dimensión temporal, aunque podría extenderse en esa dirección.

8.3. Perspectivas Futuras

Esta línea de trabajo abre múltiples vías de investigación futura:

- **Integración con variables ambientales:** Combinar embeddings de especies con datos climáticos, topográficos y de uso del suelo.
- **Modelos dinámicos:** Incorporar la dimensión temporal para estudiar cambios en distribuciones y comunidades.
- **Arquitecturas más avanzadas:** Explorar modelos transformer adaptados a datos de biodiversidad.
- **Validación ecológica:** Comparación sistemática con conocimiento ecológico establecido y validación mediante trabajo de campo.
- **Aplicaciones predictivas:** Desarrollo de modelos de nicho ecológico basados en embeddings.
- **Integración multimodal:** Combinar datos de ocurrencias con imágenes (remote sensing), datos climáticos y rasgos funcionales en un único marco de aprendizaje.

9. Conclusiones

Este trabajo ha demostrado la viabilidad y utilidad de aplicar técnicas de procesamiento del lenguaje natural al análisis de datos de biodiversidad. La analogía fundamental —especies como palabras, comunidades como frases, ecosistemas como documentos— no es meramente metafórica, sino que constituye una base sólida para aplicar décadas de avances en NLP al dominio ecológico.

GBIF proporciona el corpus ideal para este enfoque: un conjunto masivo, estandarizado y accesible de datos de biodiversidad global. La vectorización de estos datos mediante modelos Word2Vec permite operaciones matemáticas sofisticadas sobre entidades biológicas, preservando relaciones ecológicas y biogeográficas en un espacio vectorial de alta dimensión.

Las aplicaciones prácticas demostradas —detección de anomalías, predicción de hábitats, identificación de áreas submuestreadas y visualización de patrones biogeográficos— tienen relevancia directa tanto para la investigación ecológica como para la gestión y conservación de la biodiversidad.

Más allá de las aplicaciones específicas, este enfoque representa un cambio de paradigma en cómo pensamos sobre los datos de biodiversidad. Tradicionalmente, las especies se han tratado como entidades categóricas en análisis ecológicos. La vectorización semántica permite tratarlas como entidades numéricas ricas en información contextual, capaces de participar en operaciones algebraicas que reflejan relaciones ecológicas complejas.

En el contexto más amplio de la revolución de la inteligencia artificial que vivimos, este trabajo ilustra cómo técnicas desarrolladas en un dominio (el procesamiento del lenguaje) pueden transferirse creativamente a otros (la ecología y la biodiversidad). No se trata de aplicar la IA por aplicarla, sino de reconocer patrones metodológicos transferibles y adaptarlos inteligentemente a nuevos contextos.

Nuestra relación con la inteligencia artificial está en constante evolución. Como se mencionó en la introducción, podemos haber pasado por fases de asombro, escepticismo y normalización. Pero el desafío real no es ignorar estas herramientas, sino aprender a utilizarlas de manera crítica, creativa y productiva. Estamos embarcados, queramos o no, en una transformación tecnológica profunda. La cuestión no es si participar, sino cómo hacerlo de manera que maximice el beneficio para la ciencia, la conservación y la sociedad. Nuestra relación con la IA y, en concreto, con los LLM no debe quedarse la interacción en forma de chat, de buscador o de sabelotodo; los LLM se apoyan en desarrollos científicos y metodológicos que podemos usar fuera del ámbito de los propios LLM (Bengoa, 2025).

El código desarrollado y presentado en estas jornadas está disponible para cualquiera que se interese por este asunto, más como idea seminal que como producto o herramienta que también lo es. La colaboración entre gestores de datos, investigadores académicos y desarrolladores técnicos —ejemplificada en este trabajo— será esencial para aprovechar plenamente el potencial de estas nuevas metodologías.

En última instancia, este trabajo es una invitación a *"cocinar con GBIF"*: a experimentar, innovar y descubrir nuevas formas de extraer conocimiento de los vastos repositorios de datos de biodiversidad que hemos construido colectivamente. Las herramientas están disponibles, los datos están accesibles, y las posibilidades están por explorar.

Referencias

Anthropic. (2024). *Model Context Protocol Documentation*. https://modelcontextprotocol.io

Bengoa, J. (2025). *Ejemplos de aplicación de los modelos del lenguaje (LLM) en gestión forestal: gestión de la documentación y creación de herramientas de procesado. 9º Congreso Forestal Español.*

GBIF.org. (2025). *GBIF Home Page*. https://www.gbif.org

GBIF.org. (2025). *GBIF API Documentation*. https://techdocs.gbif.org/en/openapi/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. https://openai.com/blog/chatgpt

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2024). *pygbif: Python client for the GBIF API*. https://github.com/gbif/pygbif

Agradecimientos

Este trabajo ha sido posible gracias a la infraestructura proporcionada por GBIF y a la comunidad global de contribuidores de datos de biodiversidad. También queremos reconocer el papel de las herramientas de IA (Claude de Anthropic y ChatGPT de OpenAI) en el desarrollo del código y la estructuración de ideas, ejemplificando precisamente el tipo de colaboración humano-IA que este trabajo defiende.

Finalmente, agradecemos a los organizadores de las Jornadas sobre Información de Biodiversidad y Administraciones Ambientales 2025 por proporcionar el foro apropiado para presentar y discutir este trabajo.