Biblios
ISSN 1562-4730 (online)

# Estrategia de integración de métricas de impacto de la producción académica institucional mediante un Data Warehouse: caso de uso con OpenAlex, OpenAIRE y COAR

## Integration strategy for impact metrics of institutional academic output through a Data Warehouse: a case study with OpenAlex, OpenAIRE, and COAR

## Estratégia de integração de métricas de impacto da produção acadêmica institucional por meio de um Data Warehouse: estudo de caso com OpenAlex, OpenAIRE e COAR

**Pablo César de Albuquerque**
Universidad Nacional de La Plata, Buenos Aires, Argentina

**Gonzalo Luján Villarreal**
Universidad Nacional de La Plata, Buenos Aires, Argentina

ORIGINAL

## Resumen

**Objetivo.** Este artículo propone una estrategia para integrar datos de múltiples fuentes sobre producción académica, facilitando la toma de decisiones. Su enfoque es adaptable a diversas organizaciones sin restricciones en la cantidad o variedad de fuentes. **Método.** Se diseñó un sistema de integración basado en herramientas de código abierto y un modelo de datos híbrido escalable. Se combinan Data Warehouse (Kimball & Ross) para optimizar el análisis y Data Vault 2.0 para gestionar la heterogeneidad y trazabilidad, asegurando una integración flexible. **Resultados.** Se integraron los datos de OpenAIRE, OpenAlex y COAR en una tabla unificada de publicaciones académicas, combinando métricas clave como citas, vistas y descargas. La tabla incluye información relevante, como título, DOI, tipo y año de publicación, además del estado de acceso abierto. **Conclusiones.** La integración de datos permite obtener una visión más completa del impacto de la producción científica institucional. Este enfoque favorece la implementación de métricas responsables.

**Palabras clave:** producción académica institucional, identificadores persistentes, métricas responsables, Data Warehouse, Data Vault

## Abstract

**Objective.** This article proposes a strategy for integrating data from multiple sources on academic output, facilitating informed decision-making. The approach is adaptable to various organizations, regardless of the number or type of sources involved. **Method.** An integration system was designed based on open-source tools and a scalable hybrid data model. It combines Data Warehouse techniques (Kimball & Ross) to optimize analysis, and Data Vault 2.0 to manage heterogeneity and ensure traceability, enabling flexible integration. **Results.** Data from OpenAIRE, OpenAlex, and COAR were integrated into a unified academic publications table, consolidating key metrics such as citations, views, and downloads. The table includes relevant information such as title, DOI, publication type and year, as well as open access status. **Conclusions.** Data integration enables a more comprehensive view of the impact of institutional scientific output. This approach supports the implementation of responsible metrics.

**Keywords:** institutional academic output, persistent identifiers, responsible metrics, Data Warehouse, Data Vault

Resumo

**Objetivo.** Este artigo propõe uma estratégia para integrar dados de múltiplas fontes sobre a produção acadêmica, facilitando a tomada de decisões. A abordagem é adaptável a diferentes organizações, independentemente da quantidade ou tipo de fontes envolvidas. **Método.** Foi desenvolvido um sistema de integração baseado em ferramentas de código aberto e em um modelo de dados híbrido e escalável. Combina técnicas de Data Warehouse (Kimball & Ross) para otimizar a análise, e Data Vault 2.0 para gerenciar a heterogeneidade e garantir a rastreabilidade, possibilitando uma integração flexível. **Resultados.** Os dados do OpenAIRE, OpenAlex e COAR foram integrados em uma tabela unificada de publicações acadêmicas, reunindo métricas-chave como citações, visualizações e downloads. A tabela inclui informações relevantes como título, DOI, tipo e ano de publicação, além do status de acesso aberto. **Conclusões.** A integração de dados permite uma visão mais completa do impacto da produção científica institucional. Essa abordagem favorece a adoção de métricas responsáveis.

**Palavras-chave**: produção acadêmica institucional, identificadores persistentes, métricas responsáveis, Data Warehouse, Data Vault

# 1 Introducción

La evaluación del rendimiento de las instituciones académicas y científicas puede ser una herramienta útil para identificar puntos fuertes y débiles sobre el trabajo precedente, y ofrecer grandes beneficios en la toma de decisiones hacia el futuro. Una correcta evaluación debe considerar indicadores y métricas sobre el volumen de la producción científica y académica, el tamaño de la institución y el impacto que ha logrado en la comunidad. Sin embargo, dada la multiplicidad y la heterogeneidad de fuentes que brindan datos sobre producción e impacto, una evaluación que considere la producción de todos los miembros de la institución puede ser un verdadero desafío, que se complejiza a medida que el volumen de producción de la institución crece. A esto se suma la necesidad de medir actividades científicas emergentes que van más allá de la publicación de artículos científicos en revistas de alto impacto y la demanda, en crecimiento, de una evaluación más justa y transparente (Cabezas-Clavijo & Torres-Salinas, 2021).

## 1.1 Objetivo

El objetivo de este artículo es proponer una metodología que permita combinar datos sobre producción académica y científica, obtenidos de múltiples fuentes, y presentarlos de manera tal que faciliten la elaboración de tableros de control que asistan en la toma de decisiones. A fin de ejemplificar esta metodología, en este trabajo se han establecido algunas fuentes de datos y filtros iniciales para una institución específica; sin embargo, el método propuesto no supone limitaciones en cuanto a la variedad ni cantidad de fuentes de datos a utilizar, y puede adaptarse a las características y requerimientos de múltiples organismos académicos y científicos.

# 2 Revisión de literatura

En el contexto de este trabajo, la bibliometría surge como una disciplina que se ocupa de medir y analizar la producción científica a través de indicadores cuantitativos, como el número de publicaciones, las temáticas investigadas, y las instituciones y autores más prolíficos o citados, para convertirse en una herramienta imprescindible para medir la producción científica (Öztürk et al., 2024). El análisis bibliométrico permite identificar producciones realizadas por investigadores afiliados a una institución y cuantificar el impacto que genera la producción científica académica y estudiar cómo un campo de investigación evoluciona a lo largo del tiempo, lo que posibilita una mejor comprensión del desarrollo de la literatura en un área específica (Donthu et al., 2021).

En particular, la bibliometría aplicada a la producción institucional permite medir la capacidad de las instituciones para generar conocimiento, identificar áreas de especialización y optimizar sus estrategias de investigación. Estos datos son fundamentales para la autoevaluación, el reconocimiento en rankings internacionales, atraer financiamiento y fomentar colaboraciones científicas.

Para que un análisis bibliométrico sea confiable, es necesario que las métricas empleadas, sean relevantes en el contexto que se utiliza. Las métricas responsables son herramientas de evaluación que, a diferencia de las métricas tradicionales centradas únicamente en la visibilidad o el impacto inmediato, buscan reflejar de forma justa

e inclusiva tanto aspectos cuantitativos como cualitativos de la producción científica, considerando la diversidad de enfoques, metodologías y realidades regionales (Cuartas et al., 2019). Esta aproximación multidimensional integra elementos como la relevancia social, la participación comunitaria y la transferencia del conocimiento, respondiendo a la complejidad de evaluar el desempeño institucional en un contexto de proliferación de fuentes de datos y actividades emergentes. Como consecuencia, se han creado departamentos especializados en evaluación científica en diversas instituciones (por ejemplo, en la Universidad de Viena, la Universidad de Nueva Gales del Sur, la Universidad Técnica de Munich, la Universidad San Ignacio de Loyola y el Centro para el Estudio de la Ciencia y la Tecnología de la Universidad de Leiden), junto con consultoras en analítica e indicadores como Science-Metrix y el Instituto para la Información Científica de Clarivate Analytics (Cabezas-Clavijo & Torres-Salinas, 2021).

Varios autores destacan la importancia de emplear múltiples fuentes de datos en los análisis bibliométricos. Por ejemplo, Öztürk et al. (2024, p. 9) señalan que los revisores recomiendan realizar búsquedas en diversas bases de datos, mientras que Cabezas-Clavijo e Torres-Salinas (2021, p. 5) insisten en que es preferible utilizar una variedad de fuentes en lugar de depender de una sola, ya que esta podría ofrecer resultados significativos únicamente en disciplinas limitadas. En consecuencia, es fundamental seleccionar cuidadosamente las fuentes a incluir en un análisis bibliométrico para asegurar que los resultados sean tanto significativos como representativos del estado actual de la literatura en el área de estudio.

El uso de múltiples fuentes trae consigo algunas dificultades como la desambiguación de entidades, la detección de duplicados, la diferencia en la calidad de los metadatos, la evolución y cambios en los esquemas de datos. Por ejemplo, Harder (2024) describe cómo los cambios en la nomenclatura y estructura de OpenAlex impactaron en la recuperación de datos. Inicialmente, la plataforma utilizaba el término 'venue' para referirse a revistas y otros tipos de publicaciones, pero posteriormente lo reemplazó por 'source', lo que requirió modificaciones en el código para reflejar esta actualización. Además, se registraron cambios en los identificadores de autores e instituciones, lo que obligó a los investigadores a descargar nuevamente un conjunto completo de registros para garantizar la coherencia y consistencia de los datos.

En este sentido, para identificar publicaciones, autores y organizaciones de forma confiable existen los identificadores persistentes. Los identificadores persistentes son URL asignadas a recursos digitales que garantizan su localización y referencia a lo largo del tiempo. El uso de identificadores de autores no es una práctica obligatoria aunque es recomendable, ya que esto favorece la interoperabilidad entre plataformas (Albuquerque et al., 2022). Estudios como el de Aghassibake et al. (2023) demuestran que, aunque herramientas basadas en identificadores persistentes (PID, por sus siglas en inglés), como Open Researcher and Contributor ID (ORCID) y Digital Object Identifier (DOI), facilitan el análisis de colaboraciones internacionales y la generación de visualizaciones, también evidencian desafíos en la calidad de los metadatos, como la falta de asociación de autores a un ORCID o perfiles desactualizados.

Para abordar estos desafíos de integración, el enfoque propuesto en este artículo aprovecha diversas fuentes de datos de alta calidad, como *Science Knowledge Graphs* (SKG), repositorios institucionales y vocabularios controlados desarrollados por expertos como los de Confederation of Open Access Repositories (COAR). En particular, se utilizan dos SKG, OpenAIRE Graph (Manghi et al., 2019) y OpenAlex (Priem et al., 2022), que permiten representar entidades científicas y sus relaciones de forma procesable (Hogan et al., 2021; Ciuciu-Kiss & Garijo, 2024). La integración de estas fuentes presenta desafíos derivados de las diferencias en la organización, calidad y formatos de los metadatos, lo que requiere mecanismos eficientes de desambiguación y normalización. Mientras OpenAIRE Graph consolida datos de repositorios institucionales, sistemas Current Research Information System (CRIS) y editores, OpenAlex (sucesor del Microsoft Academic Graph) conecta obras, autores, instituciones, conceptos y fuentes mediante identificadores persistentes y ofrece una Application Programming Interface (API) intuitiva para el acceso a sus datos.

La incorporación de vocabularios controlados, como el *Resource Type Vocabulary* (RTV) de COAR, no sólo facilita la interoperabilidad y clasificación estandarizada de los recursos científicos a nivel global, sino que también mejora la recuperación de información, permite la normalización de tipologías documentales y contribuye a una mejor integración con infraestructuras de acceso abierto y herramientas de análisis avanzadas. Los metadatos de las publicaciones pueden tener diferentes términos para referirse a un mismo tipo de documento (ej. "*preprint*", "*manuscript*", "*working paper*"). El RTV de COAR proporciona un conjunto unificado de categorías que permite homogeneizar estas denominaciones, asegurando que documentos similares sean reconocidos como tales, independientemente de la fuente de origen.

# 3 Metodología

Para integrar datos sobre producción científica a partir de múltiples fuentes, se ha diseñado un sistema cuyo objetivo es la creación de una tabla consolidada de publicaciones. Para ello, se emplean herramientas de código abierto y un modelo de datos híbrido que combina enfoques dimensionales, organizando la información en hechos y dimensiones, tal como proponen Kimball & Ross (2013) en su metodología para el desarrollo de *Data Warehouse*. Un *Data Warehouse* es un sistema de almacenamiento de datos diseñado para facilitar el análisis y la toma de decisiones mediante la consolidación y organización de grandes volúmenes de datos provenientes de diversas fuentes. Estas tablas de hechos y dimensiones se implementan en base a la metodología *Data Vault 2.0* (Linstedt & Olschimke, 2015), lo que permite gestionar la heterogeneidad y garantizar la consistencia de los datos.

La estrategia adoptada sigue un enfoque Extract, Load, Transform, (ELT) en el que los datos son extraídos desde las fuentes y almacenados sin modificaciones en una base de datos relacional, conservando su estructura original. Este enfoque, que surge como evolución del modelo Extract, Transform, Load (ETL) tradicional ante la llegada del *Big Data*, permite manejar grandes volúmenes de información al aprovechar la potencia de cómputo distribuido y la escalabilidad inherente a los almacenes de datos modernos (Dhaouadi et al., 2022).

La recolección inicial de datos se lleva a cabo mediante el uso de interfaces de programación de aplicaciones (APIs), específicamente la API de OpenAlex y la API del OpenAIRE Graph. Estas APIs permiten acceder de forma estructurada y eficiente a datos académicos y científicos distribuidos globalmente, facilitando su extracción directa mediante PID institucionales como el Research Organization Registry (ROR). Asimismo, para la estandarización y clasificación consistente de los tipos de recursos obtenidos, se utiliza el vocabulario controlado de COAR.

A diferencia del ETL, que realiza tareas de limpieza y homogeneización antes de la carga, el ELT posterga estas transformaciones hasta después del almacenamiento inicial. Esta característica resulta particularmente beneficiosa en el contexto de datos académicos, cuya naturaleza distribuida y crecimiento constante en métricas de impacto requiere una capacidad continua de monitoreo y análisis retrospectivo. El enfoque ELT permite gestionar eficientemente estos datos dinámicos, asegurando que información considerada poco relevante en un momento dado pueda ser reevaluada en etapas posteriores, respondiendo así a cambios en las tendencias de investigación o criterios de evaluación científica.

Adicionalmente, esta metodología proporciona una mayor trazabilidad de los datos, aspecto crucial en la evaluación institucional de la ciencia, dado que permite registrar y auditar todas las transformaciones realizadas, desde su forma original hasta los modelos dimensionales específicos para cada propósito de análisis. En este sentido, ELT se posiciona como una estrategia adecuada para entornos que demandan procesamiento sobre grandes conjuntos de datos académicos y requieren accesibilidad y análisis casi inmediatos, superando posibles cuellos de botella de escalamiento que se presentan típicamente en enfoques ETL tradicionales.

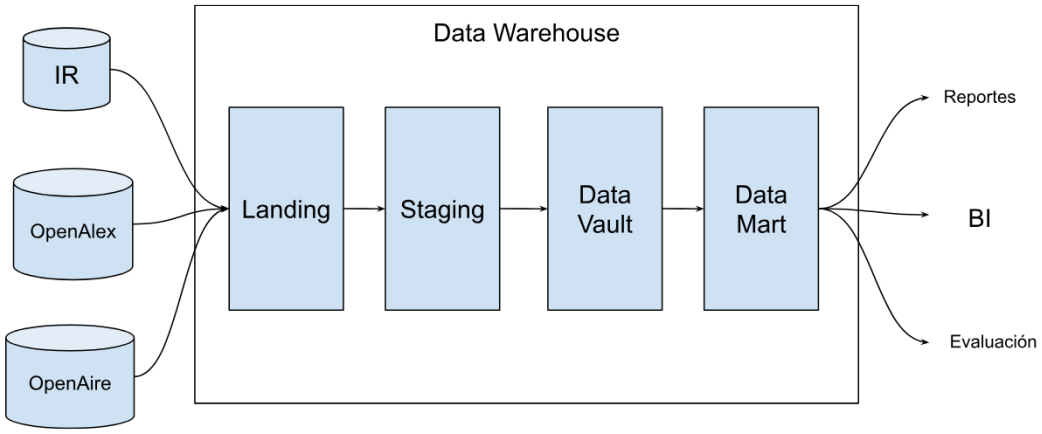
El uso de modelos de *Data Warehouse* en el ámbito académico ha demostrado ser una estrategia efectiva para consolidar información dispersa y facilitar el análisis de la producción científica. En trabajos previos, se ha propuesto la construcción de un *Data Warehouse* a partir de los datos disponibles en los repositorios institucionales, permitiendo integrar y unificar la información con el fin de optimizar la toma de decisiones en instituciones académicas (Albuquerque et al., 2021). Asimismo, se ha explorado un enfoque dimensional basado en la metodología de Kimball como una solución adaptable a distintas fuentes de datos, independientemente de su origen, lo que facilita la comparación y el seguimiento de la producción académica a lo largo del tiempo (Albuquerque et al., 2023). Casos recientes también han aplicado con éxito procesos ETL/ELT para integrar datos académicos y científicos en contextos institucionales a nivel nacional, demostrando la efectividad de estas técnicas en la creación de fuentes únicas de información para el análisis bibliométrico y la toma de decisiones estratégicas (Silva et al., 2022; Tomczyńska et al., 2023). Estos antecedentes refuerzan la aplicabilidad del modelo propuesto en este trabajo, destacando su flexibilidad y capacidad para adaptarse a diferentes entornos institucionales y sistemas de información.

Para optimizar la legibilidad y la estructura narrativa del trabajo, se utilizó la asistencia de la inteligencia artificial (IA) Gemini 1.5 Flash. El proceso consistió en proporcionar a la herramienta fragmentos de texto extensos junto con contexto específico sobre la disciplina (Ciencias de la Información y Bibliometría). Se utilizaron prompts tales como: 'Ayúdame a desarrollar este bloque de texto en dos o tres párrafos más breves para favorecer la narrativa y la claridad, considerando que el tema es la integración de métricas mediante Data Warehouse'. Tras las sugerencias de segmentación y redacción propuestas por la IA, los autores realizaron una revisión crítica para asegurar la precisión técnica y el rigor científico del contenido final.

## 3.1 Arquitectura

La arquitectura del sistema propuesto se organiza en cuatro capas: *Landing*, *Staging*, *Data Vault* y *Data Mart*. Cada capa cumple un rol específico en el procesamiento de un conjunto de datos, desde la extracción hasta la presentación final. La Figura 1 ilustra este flujo, detallando cómo los datos son ingeridos desde las fuentes, que incluyen OpenAlex, OpenAIRE y repositorios institucionales (IR, por sus siglas en inglés), procesados secuencialmente a través de cada una de las capas y, finalmente, expuestos para su consumo en forma de Reportes, Business Intelligence (BI) y Evaluación.

**Figura 1**
*Arquitectura del Data Warehouse*




***Nota.*** Fuente: Elaboración propia.

### 3.1.1 Landing

En la capa *Landing*, los datos se almacenan tal como se extraen de las fuentes, sin aplicar transformaciones ni modificaciones. Esto permite preservar su esquema original y delegar las transformaciones a etapas posteriores, asegurando la conservación de los datos sin alteraciones. Este enfoque resulta especialmente útil para tareas de depuración, donde es necesario revisar los datos originales ingestados. Además, los procesos de extracción en esta etapa están diseñados para minimizar el impacto en los sistemas de origen, regulando el uso de recursos, especialmente en fuentes con cuotas o límites de descarga.

### 3.1.2 Staging

La capa *Staging* inicia el procesamiento de los datos, realizando las primeras tareas de limpieza y normalización. Estas incluyen la eliminación de duplicados, la imputación de valores faltantes, la definición de tipos de datos para cada atributo y la estandarización de nombres según una convención propia.

En esta etapa también se generan claves primarias para identificar de manera unificada las entidades recuperadas, sin depender de sus claves de negocio (*business keys*). Estas últimas son identificadores relevantes para nuestra investigación, pero pueden variar en su representación según la fuente de origen.

Las claves primarias se generan a partir de datos ya normalizados, lo que permite compararlas con claves equivalentes de otras fuentes también normalizadas. Una normalización precisa es fundamental, ya que cualquier error en este proceso podría generar discrepancias en la comparación de claves primarias de una misma entidad, afectando la integración de los datos.

Por ejemplo, OpenAlex representa los DOI con la URL completa, mientras que OpenAIRE los almacena en un formato diferente. Para generar claves primarias comparables, es necesario normalizar los DOI y unificarlos en un mismo formato, como conservar únicamente el prefijo y el sufijo del identificador.

### 3.1.3 Data Vault

En la tercera capa, *Data Vault*, se modelan los datos procesados previamente siguiendo la metodología *Data Vault 2.0*, dividiendo la información en tres componentes fundamentales:

a) ***Hubs*:** Almacenan las claves de negocio, por ejemplo, DOI para publicaciones u ORCID para autores. Se incluyen metadatos adicionales como la fuente y la fecha de carga, lo que permite rastrear el origen de los datos.

b) ***Links***: Representan las relaciones entre entidades, conectando dos o más *Hubs*. Estos *Links* contienen las claves de los *Hubs* relacionados y algunos metadatos para auditar el origen de la relación, sin agregar información temporal o contextual adicional.

c) ***Satellites*:** Guardan los atributos descriptivos y contextuales de las entidades o relaciones, permitiendo registrar cambios a lo largo del tiempo. Por ejemplo, un *Satellite* asociado a un DOI puede almacenar el título, el año de publicación y la cantidad de citas, ofreciendo una visión dinámica del objeto de negocio.

Esta estructura posibilita mantener la integridad histórica de los datos y facilita la actualización ante cambios en las fuentes o en los requisitos del análisis.

### 3.1.4 Data Mart

La capa final, el *Data Mart*, se encarga de presentar la información de manera accesible y optimizada para el usuario final. Aquí se estructuran los datos en modelos dimensionales, compuestos por hechos y dimensiones, que facilitan consultas rápidas y análisis intuitivos. Los hechos contienen las métricas principales (como la cantidad de citas en una publicación), mientras que las dimensiones aportan el contexto (por ejemplo, autor, institución o año de publicación). Este modelo dimensional se construye a partir de la capa previa, integrando datos de múltiples fuentes mediante los *Hubs*, *Satellites* y *Links* que conectan estas entidades.

## 3.2 Estrategia de recolección de datos

Para recuperar los recursos de una institución específica, se utiliza su identificador *Research Organization Registry* (ROR). El proceso de extracción varía según la fuente de datos.

### 3.2.1 Extracción desde OpenAlex

La API de OpenAlex ofrece el endpoint */works* aplicando el filtro *institutions.id* con el ROR correspondiente (*Works | OpenAlex Technical Documentation*, 2023). Este filtro se basa en el campo *authorships.institutions.ror* (o a través del alias *institutions.ror*), lo que permite identificar únicamente las publicaciones asociadas a autores afiliados a la organización. Por ejemplo, a partir de la siguiente consulta es posible recuperar datos de la Universidad Nacional de La Plata (UNLP):

https://api.openalex.org/works?filter=institutions.ror:https://ror.org/01tjs6929

Además, se utiliza el *endpoint/authors* con el filtro *affiliations.institution.ror*, para recuperar información sobre autores. Aunque estos datos pueden recuperarse desde el objeto *authorships* de *works*, este *endpoint* proporciona información más detallada, como los años en que el autor estuvo ligado a la institución.

### 3.2.2 Extracción desde OpenAIRE Graph

En OpenAIRE, los recursos se recuperan a través del endpoint */researchProducts*, aplicando filtros en los parámetros *relOrganizationId* y *relCollectedFromDatasourceId* (*Filtering Search Results | OpenAIRE Graph Documentation*, 2025; *Searching Entities | OpenAIRE Graph Documentation*, 2025).

El filtro *relOrganizationId* permite recuperar los recursos asociados a una organización a partir de su identificador en OpenAIRE. Sin embargo, OpenAIRE no mantiene un identificador interno para autores ni una relación explícita entre estos y las instituciones. La única forma de verificar con certeza la afiliación institucional de un autor sería mediante su ORCID, pero esta información no siempre está disponible.

En cuanto al filtro *relCollectedFromDatasourceId*, este se utiliza para obtener datos de fuentes específicas dentro de la institución, como por ejemplo su repositorio institucional. Aunque esto no garantiza que los autores estén afiliados a la institución, sí ofrece un fuerte indicio de su relevancia, ya que estos repositorios suelen contener producción científica vinculada a la organización.

El identificador de OpenAIRE de una organización puede obtenerse a partir de su ROR. Por ejemplo, siendo el ROR de UNLP "https://ror.org/01tjs6929" puede obtenerse datos a partir de la siguiente consulta: https://api.openaire.eu/graph/organizations?pid=https://ror.org/01tjs6929

Es posible ver las fuentes de datos relacionadas con la UNLP en OpenAIRE accediendo al siguiente enlace. Este utiliza un identificador que representa a la universidad:

https://api.openaire.eu/graph/dataSources?relOrganizationId=openorgs____::40b9f835648a3e0d057d6917dd7e54d5. Allí se muestran los sistemas o plataformas que OpenAIRE vincula con la UNLP, entre los que se encuentra el repositorio institucional.

Una vez obtenido su identificador, se puede utilizar el *endpoint /researchProducts* para recuperar los recursos depositados en el repositorio de la siguiente manera:

https://api.openaire.eu/graph/researchProducts?relOrganizationId=openorgs____::40b9f835648a3e0d057d6917dd7e54d5

### 3.2.3 Estrategia de integración de fuentes

Una vez que los datos ya fueron normalizados y ubicados en el *Data Vault* se produce la integración de ambas fuentes en la capa de *Data Mart*.

Para integrar los datos recuperados de OpenAIRE y OpenAlex, utilizamos los DOI como identificadores de los recursos institucionales. Sin embargo, es importante considerar cómo cada fuente maneja estos identificadores.

Por ejemplo, una tabla de publicaciones puede unificar registros de OpenAlex y OpenAIRE a través de sus respectivos *Hubs* de DOI, permitiendo además la consolidación de métricas como citas, extraídas del *Satellite* de OpenAlex, y vistas y descargas, recuperadas del *Satellite* de OpenAIRE. De esta manera, los datos de ambas fuentes se estructuran de forma coherente, facilitando su análisis en el modelo dimensional.

En algunos casos, una publicación puede tener más de un DOI, por ejemplo, uno asignado a una versión *preprint* en arXiv y otro a la versión publicada. Sin embargo, OpenAlex solo almacena un único DOI por publicación, priorizando el correspondiente a la versión final (Work Object | OpenAlex Technical Documentation, 2025).

Por otro lado, OpenAIRE Graph aplica un proceso de deduplicación en el que agrupa distintos identificadores persistentes (PID), permitiendo que un mismo recurso conserve múltiples DOI.

## 3.3 Implementación

Para la implementación del proceso de recolección, integración y análisis de datos, se utilizaron dos herramientas *Open Source* que facilitaron la extracción, normalización, modelado y visualización de la información recuperada desde múltiples fuentes, dbt (Albuquerque, 2024a) y Kedro (Albuquerque, 2024b).

Kedro es un *framework* de Python diseñado para desarrollar proyectos de ciencia de datos de forma reproducible y mantenible. Un proyecto en Kedro se organiza en *pipelines*, los cuales están formados por nodos. Cada nodo es, esencialmente, una función de Python que recibe datos de entrada y genera una salida. Los *pipelines* definen cómo se conectan estos nodos, estructurando el flujo de datos de manera clara y modular.

Además, Kedro incluye un catálogo de fuentes de datos y está integrado con Jupyter Notebook, lo que facilita la exploración de datos, el perfilado y la ejecución de código experimental antes de integrarlo a los nodos (Data Catalog, 2025; Filtering Search Results | OpenAIRE Graph Documentation, 2025).

La segunda herramienta utilizada específicamente para la transformación de datos fue dbt (Data Build Tool) ya que permite estructurar, documentar y gestionar modelos analíticos dentro de un entorno Structured Query Language (SQL). En este trabajo, se empleó para procesar los datos extraídos, organizándolos inicialmente en un *Data Vault* y posteriormente en un modelo dimensional optimizado para el análisis.

### 3.3.1 Implementación de la extracción

Kedro se utiliza, en una primera instancia, para la extracción y el almacenamiento inicial de los datos en la etapa de *Landing*. Se implementó un *pipeline* por cada fuente: uno para OpenAlex y otro para OpenAIRE Graph. Cada *pipeline* consta de dos nodos principales: uno para la extracción de datos y otro para su normalización y almacenamiento en la base de datos.

Los nodos de extracción se encargan de establecer la conexión con la fuente de datos y aplicar la estrategia de recolección, ya sea completa o incremental, utilizando parámetros como fechas o identificadores persistentes (PID). Los datos obtenidos se almacenan en archivos en formato *Parquet*.

Los nodos de normalización y almacenamiento procesan estos archivos, adaptando su estructura para su inserción en la base de datos. Dado que los archivos pueden contener datos anidados dentro de una misma columna, es necesario desnormalizarlos antes de su carga. Por ejemplo, en OpenAlex, los objetos *Authorship* de las entidades *Work* requieren ser transformados para que cada autor se represente como una entrada independiente. Además, se incorporan metadatos, como la fecha de carga, para garantizar la trazabilidad en el *Data Warehouse*. Finalmente, los datos procesados se almacenan en un esquema de *Landing* dentro de la base de datos relacional.

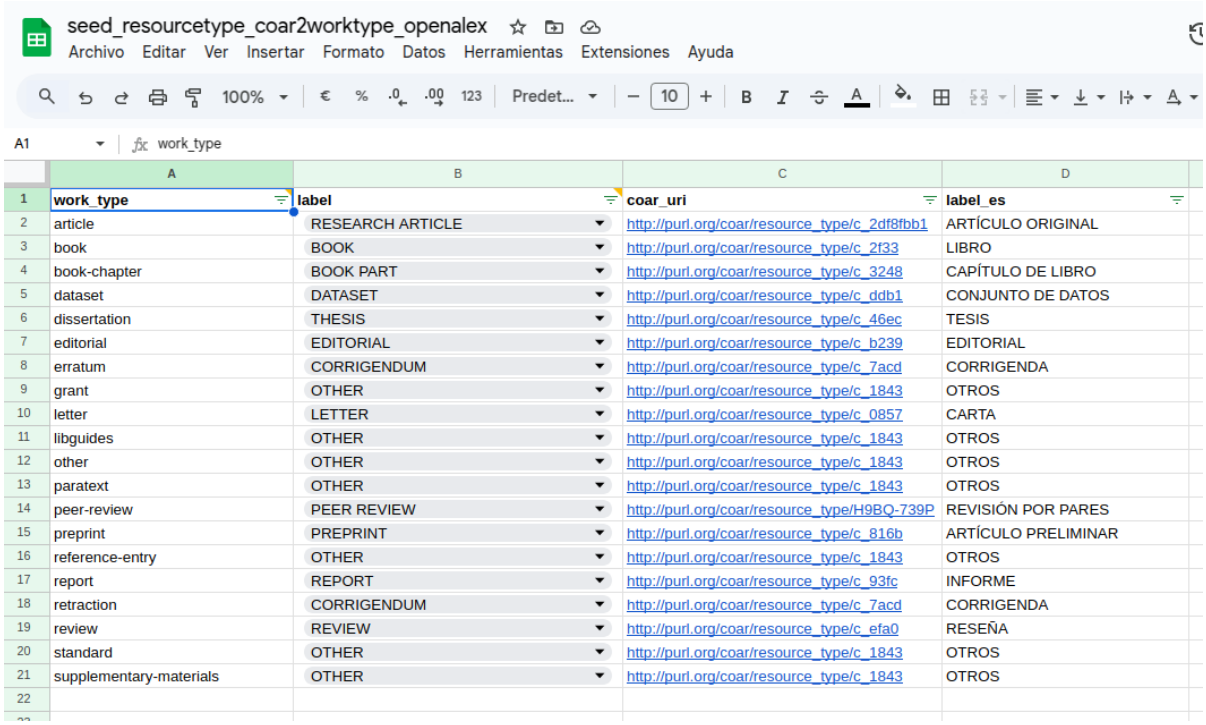
### 3.3.2 Normalización y modelado con dbt

Con dbt se realizan principalmente dos tareas, la carga de mapeos de los tipos de recursos de cada fuente con el vocabulario de COAR y luego el procesamiento de los datos normalizados en *Staging*, a partir de los datos depositados en *Landing*, que finalmente serán insertados en el *Data Vault*.

La carga de los mapeos se realiza a partir del uso de *seeds*, utilizando archivos Comma-Separated Values (CSV) para generar tablas auxiliares (Add Seeds to Your DAG | Dbt Developer Hub, 2025). Un ejemplo concreto de esto es el archivo de mapeo para los tipos de recursos. En la Figura 2 se observa la estructura de este CSV, donde se establece la correspondencia directa entre el campo *work_type* de OpenAlex y las etiquetas (en inglés y español) y URIs del vocabulario RTV de COAR.

**Figura 2**

*Ejemplo de CSV generado para mapear los tipos de datos en OpenAlex con RTV de COAR.*



seed_resourcetype_coar2worktype_openalex

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda

A1 work_type

| | A | B | C | D |
|---|---|---|---|---|
| 1 | work_type | label | coar_uri | label_es |
| 2 | article | RESEARCH ARTICLE | http://purl.org/coar/resource_type/c_2df8fbb1 | ARTÍCULO ORIGINAL |
| 3 | book | BOOK | http://purl.org/coar/resource_type/c_2f33 | LIBRO |
| 4 | book-chapter | BOOK PART | http://purl.org/coar/resource_type/c_3248 | CAPÍTULO DE LIBRO |
| 5 | dataset | DATASET | http://purl.org/coar/resource_type/c_ddb1 | CONJUNTO DE DATOS |
| 6 | dissertation | THESIS | http://purl.org/coar/resource_type/c_46ec | TESIS |
| 7 | editorial | EDITORIAL | http://purl.org/coar/resource_type/c_b239 | EDITORIAL |
| 8 | erratum | CORRIGENDUM | http://purl.org/coar/resource_type/c_7acd | CORRIGENDA |
| 9 | grant | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 10 | letter | LETTER | http://purl.org/coar/resource_type/c_0857 | CARTA |
| 11 | libguides | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 12 | other | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 13 | paratext | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 14 | peer-review | PEER REVIEW | http://purl.org/coar/resource_type/H9BQ-739P | REVISIÓN POR PARES |
| 15 | preprint | PREPRINT | http://purl.org/coar/resource_type/c_816b | ARTÍCULO PRELIMINAR |
| 16 | reference-entry | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 17 | report | REPORT | http://purl.org/coar/resource_type/c_93fc | INFORME |
| 18 | retraction | CORRIGENDUM | http://purl.org/coar/resource_type/c_7acd | CORRIGENDA |
| 19 | review | REVIEW | http://purl.org/coar/resource_type/c_efa0 | RESEÑA |
| 20 | standard | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 21 | supplementary-materials | OTHER | http://purl.org/coar/resource_type/c_1843 | OTROS |
| 22 | | | | |

***Nota.*** Fuente: Elaboración propia.

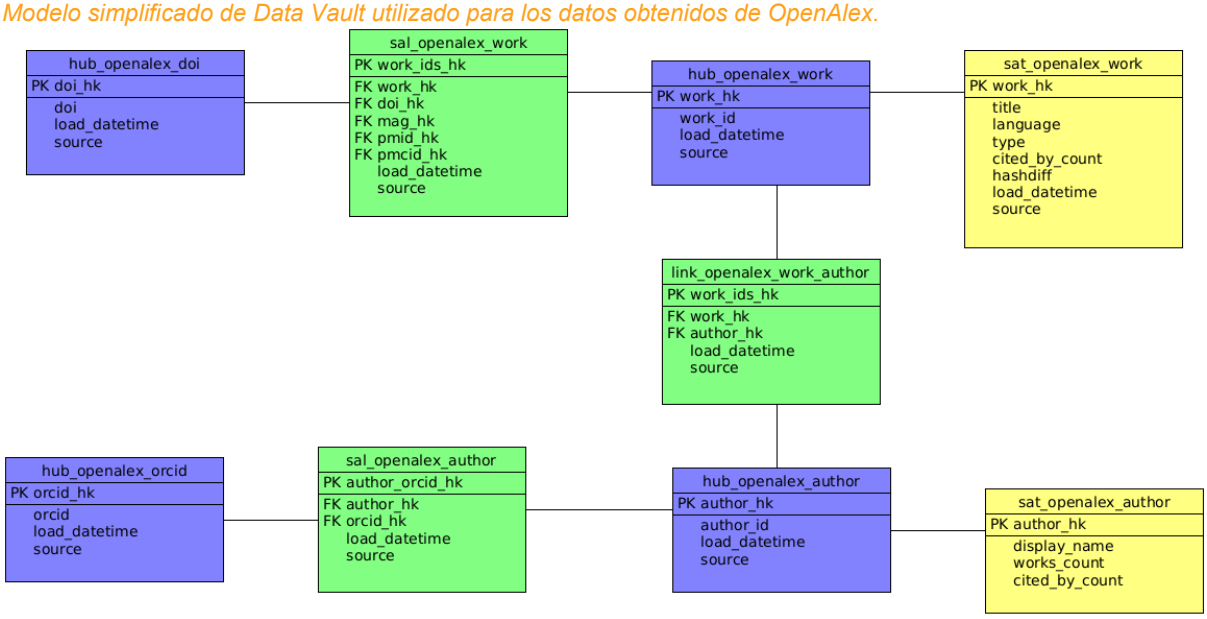
La normalización de datos incluye el filtrado de atributos irrelevantes, el renombramiento de columnas para garantizar consistencia, la estandarización de identificadores persistentes como DOI y URL, y el tratamiento de valores nulos mediante la definición de tipos de datos adecuados.

En la etapa de modelado en *Data Vault*, se crearon vistas de *Staging* donde se establecen las claves primarias definitivas para identificar de manera única cada entidad y sus relaciones, dentro del *Data Warehouse*. Posteriormente, se definieron las estructuras de *Hubs*, *Links* y *Satellites*, como se puede ver en las siguientes imágenes. Estas estructuras se implementaron de forma específica para cada fuente de datos, modelando sus entidades principales.

La Figura 3 presenta una versión simplificada del modelo *Data Vault* para OpenAlex, enfocándose en la organización en torno al *Hub* que contiene los identificadores de *Work* en esa fuente (*hub_openalex_work*) y sus datos contextuales, en su *Satellite* correspondiente (*sat_openalex_work*). El diagrama también muestra las relaciones de este *Hub* con otros identificadores, como el de autores (*hub_openalex_author*) y el de DOI (*hub_openalex_doi*).

**Figura 3**

*Modelo simplificado de Data Vault utilizado para los datos obtenidos de OpenAlex.*

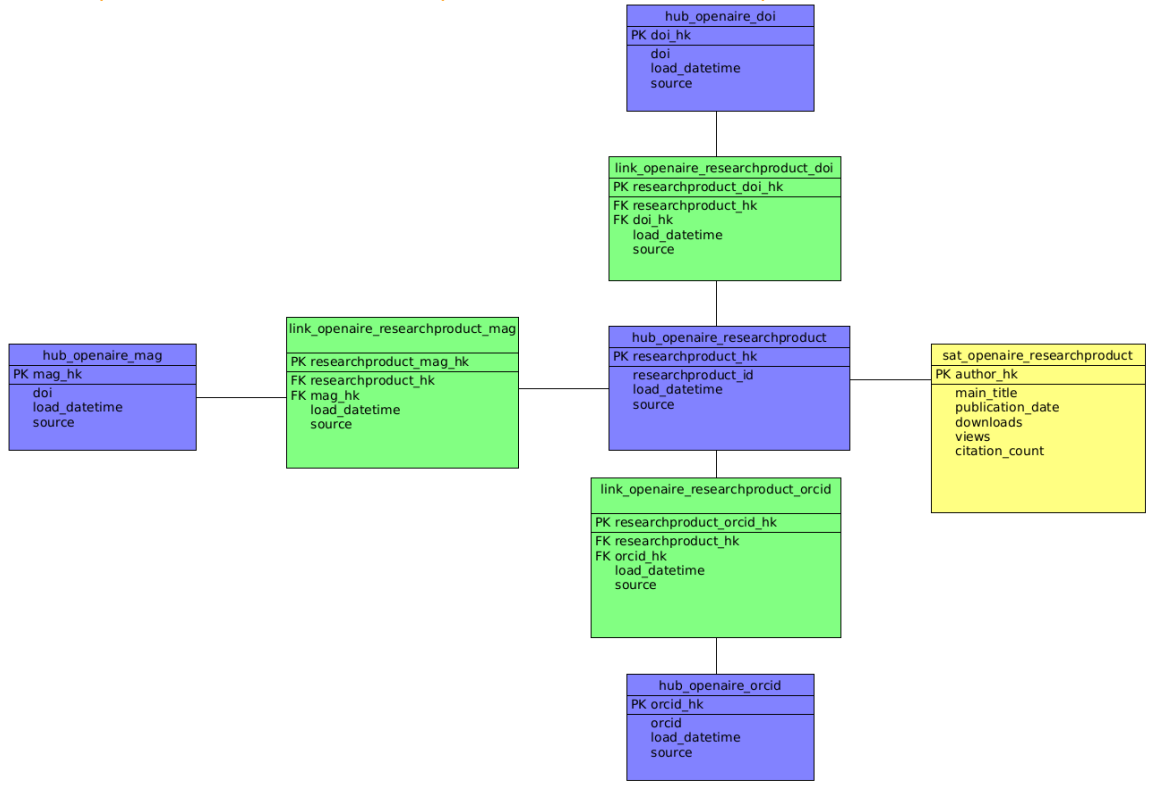



*Nota.* Fuente: Elaboración propia.

A continuación, la Figura 4 presenta el modelo simplificado correspondiente para OpenAIRE. En esta fuente, la entidad de publicación se denomina *Research Product*, por lo que el modelo se enfoca en el *Hub* que almacena los identificadores de esta entidad (*hub_openaire_researchproduct*) y su satélite de datos contextuales (*sat_openaire_researchproduct*). Adicionalmente, se muestran las relaciones con otros *Hubs* dedicados a identificadores persistentes como DOI, MAG y ORCID (*hub_openaire_doi*, *hub_openaire_mag* y *hub_openaire_orcid*).

**Figura 4**

*Modelo simplificado de Data Vault utilizado para los datos obtenidos de OpenAire.*

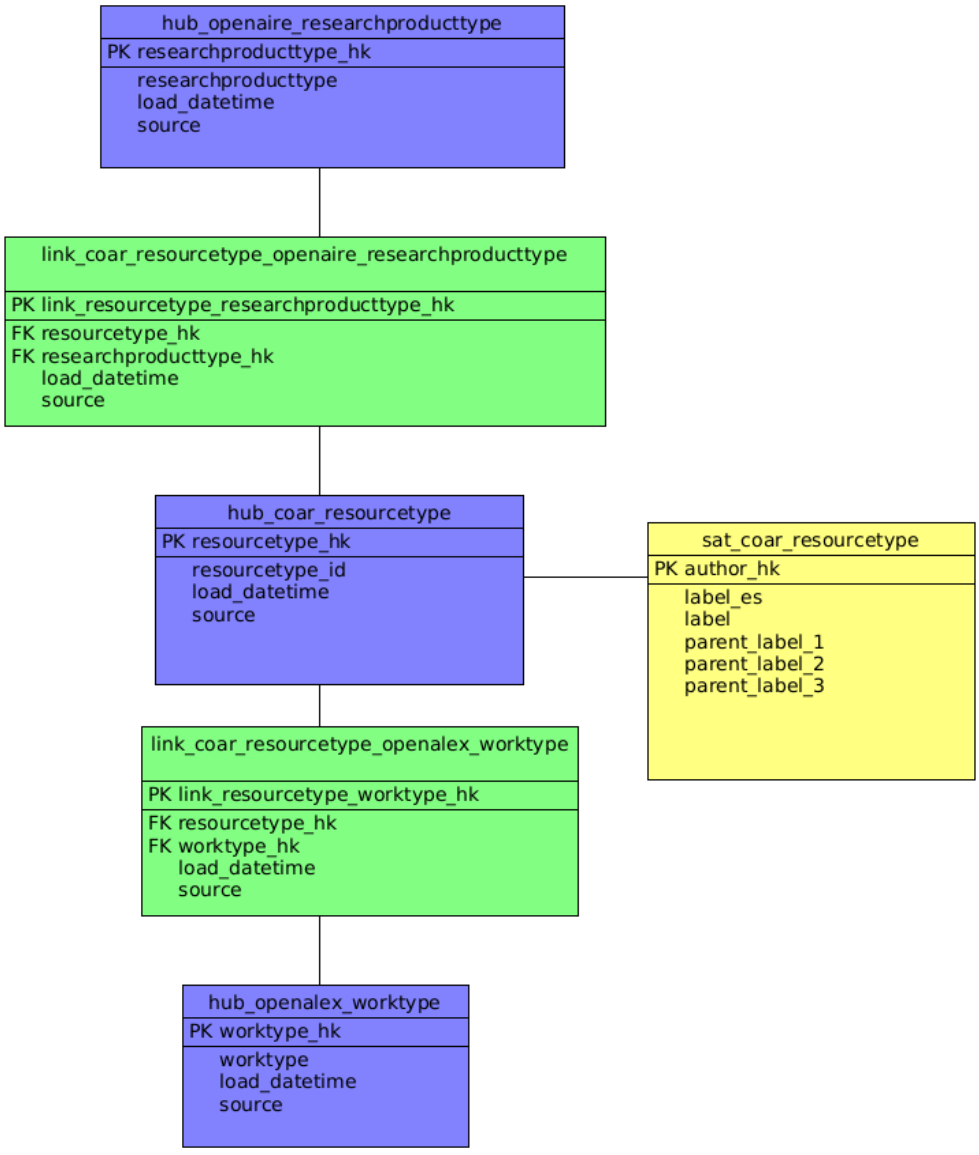



*Nota.* Fuente: Elaboración propia.

La estandarización de metadatos es un pilar fundamental en la estrategia de integración, permitiendo una visión

unificada de los datos de diversas fuentes. Para lograr esto, se implementó un modelo *Data Vault* que integra los tipos de recursos de OpenAlex y OpenAIRE utilizando el vocabulario controlado de COAR. La Figura 5 ilustra este modelo simplificado, enfocado en el *hub_coar_resourcetype*. Este *Hub* centraliza los tipos de recursos y, a través de *Links* específicos, se conecta con los *Hubs* que almacenan los tipos de documentos tal como son definidos en OpenAlex (*hub_openalex_worktype*) y OpenAIRE (*hub_openaire_researchproducttype*), así como con su satélite que contiene las descripciones de los tipos de recursos de COAR.

**Figura 5**
*Modelo simplificado de Data Vault utilizado para los datos obtenidos de COAR.*




*Nota.* Fuente: Elaboración propia.

### 3.3.3 Integración y visualización de datos unificados

La integración final se realizó mediante la intersección de los datos de OpenAIRE y OpenAlex utilizando el identificador DOI como clave común, consolidando en una tabla unificada las publicaciones institucionales junto con las métricas clave provenientes de cada fuente (cantidad de citas, vistas y descargas). Esta tabla consolidada, ubicada en la capa de *Data Mart*, fue construida a partir del modelo *Data Vault* desarrollado en la capa previa, proporcionando una estructura adecuada para la presentación y análisis de resultados.

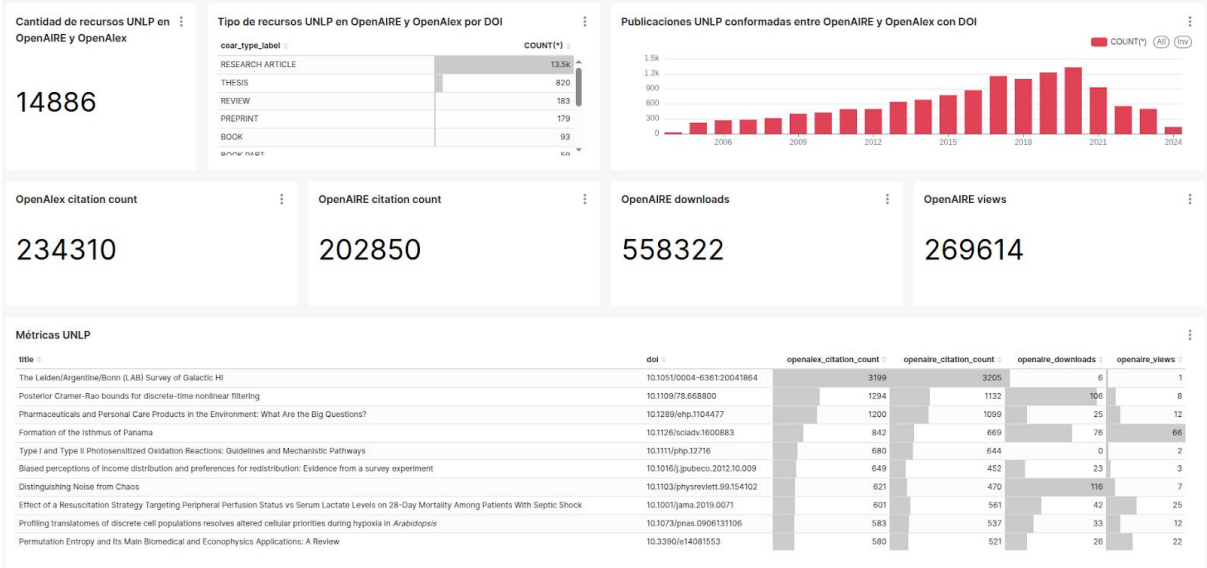
A partir de esta tabla, se implementó un *dashboard* interactivo en Apache Superset que permite explorar la producción científica institucional, comparar indicadores entre fuentes, y analizar la evolución temporal de su visibilidad e impacto.

## 4 Resultados

Para obtener el conjunto de datos final se vincularon mediante una intersección entre las tablas procesadas provenientes de OpenAIRE y OpenAlex. Esta integración permitió mantener únicamente las publicaciones claramente identificables como producción científica institucional, excluyendo recursos cuya filiación institucional no pudiera establecerse con certeza. Un ejemplo es el caso Philosophiæ naturalis principia mathematica, de Isaac Newton. Si bien se trata de una obra con alto valor patrimonial, su inclusión podría distorsionar el análisis del impacto de la producción científica contemporánea de la institución.

La Figura 6 muestra un reporte interactivo que integra datos provenientes de OpenAlex, OpenAIRE y COAR. Fue desarrollado con una estética y disposición de gráficos similar a la que utiliza OpenAlex para visualizar información institucional, como se observa en el caso de la UNLP (OpenAlex | Universidad Nacional de La Plata, 2025). El tablero presenta un contador con el total de publicaciones institucionales, un gráfico de barras que muestra la evolución anual de publicaciones, y una tabla que lista las obras de autores institucionales ordenadas según la cantidad de citas registradas por OpenAlex. Además, se incluye información sobre vistas y descargas provistas por OpenAIRE, así como otra tabla que clasifica las publicaciones según el vocabulario de tipos de recursos definido por COAR. Finalmente, se destacan las métricas agregadas de citas totales según OpenAlex, y de vistas, descargas y citas contabilizadas por OpenAIRE.

**Figura 6**
*Datos de la UNLP en un dashboard desarrollado con Apache Superset*




*Nota.* Fuente: Elaboración propia.

Desde OpenAIRE se recuperaron 159.807 recursos depositados en el repositorio institucional SEDICI, de los cuales 74.866 tienen DOI asignado. Por su parte, en OpenAlex se identificaron 39.564 publicaciones asociadas a autores con filiación explícita a la Universidad Nacional de La Plata (UNLP). Para este conjunto se registraron 39.570 DOI, debido a que algunas publicaciones pueden estar asociadas a más de un identificador.

La integración de ambas fuentes, utilizando exclusivamente DOI como clave común debido a su amplia adopción, permitió identificar 14.886 publicaciones coincidentes. Si bien este número podría incrementarse incorporando otros identificadores persistentes (PID), como MAG o PMID, se optó por simplificar el análisis usando únicamente DOI.

Las métricas consolidadas muestran 234.310 citas registradas en OpenAlex, y 202.850 citas, 558.322 descargas y 269.614 vistas según OpenAIRE. Adicionalmente, las publicaciones institucionales fueron clasificadas utilizando el vocabulario COAR, destacando 13.481 artículos de investigación, 820 tesis, 183 revisiones, 179 preprints, 93 libros, 59 capítulos de libros, además de otros tipos documentales minoritarios.

## 5 Análisis y discusión

Los resultados destacan cómo la combinación de OpenAlex y OpenAIRE contribuye a generar un *dataset* preciso y trazable, en el que se prioriza la identificación inequívoca de las publicaciones institucionales mediante DOI y la verificación de la afiliación a partir del uso de ROR. OpenAlex aporta precisión al identificar publicaciones con autores claramente afiliados a la institución, mientras que OpenAIRE complementa esta información con valiosas métricas de citas, vistas y descargas. Esta integración no solo incrementa la calidad del conjunto de datos resultante, sino que también proporciona información estratégica útil para administradores de repositorios institucionales interesados en evaluar el impacto de sus publicaciones en OpenAlex y comparar diferentes fuentes para obtener un conteo más preciso de citas.

La integración propuesta, que utiliza un modelo híbrido *Data Warehouse* y *Data Vault 2.0*, demostró eficacia en la gestión de datos heterogéneos y complejos, alineándose con los objetivos del estudio al permitir estructurar coherentemente datos provenientes de diversas fuentes para facilitar su análisis y visualización.

En comparación con otros trabajos previos como el estudio de Silva et al. (2022), que implementan una estrategia ETL (*Extraction, Transformation, Load*) para integrar plataformas como LA Referencia y VIVO, el enfoque presentado aquí confirma y amplía la relevancia de estas estrategias robustas al incorporar específicamente la metodología *Data Vault 2.0*. Este modelo ofrece flexibilidad y precisión adicional, contribuyendo significativamente al campo de la bibliometría institucional y a la gestión de repositorios digitales al permitir un análisis más completo y riguroso de la producción científica.

Sería incluso interesante considerar la integración directa de los propios repositorios institucionales como fuentes de datos. Esto permitiría incorporar registros que no hayan sido recolectados por OpenAIRE o corregir posibles errores introducidos durante su procesamiento, como normalizaciones incorrectas de metadatos o registros fusionados erróneamente. Para ello, se podría acceder directamente a la base de datos del repositorio, si estuviera disponible, o bien emplear estrategias externas como OAI-PMH, API REST o incluso *web scraping*.

Al estructurar los datos en un formato adecuado y visualizarlos mediante herramientas como Apache Superset, se facilita la identificación de patrones, tendencias y áreas de colaboración dentro de la institución. Esto no solo contribuye a la evaluación del impacto académico, sino que también respalda la toma de decisiones estratégicas en investigación y desarrollo.

Si bien existen diversas herramientas de visualización de datos, tanto en la nube (como Looker Studio o Tableau) como de escritorio (como Power BI), la elección de Apache Superset ofrece un equilibrio entre flexibilidad y control. Permite funcionalidades propias de las soluciones en la nube, como la posibilidad de compartir tableros de control y trabajar colaborativamente, pero sin depender de un servicio externo. Además, al estar licenciado bajo Apache 2.0, proporciona mayor libertad para su personalización e integración con otros sistemas, garantizando un entorno adaptable a las necesidades institucionales.

En trabajos futuros, se integrarán nuevas entidades en el análisis, como los tópicos (en OpenAlex) o materias (subjects en OpenAIRE), que proporcionan información adicional sobre las áreas temáticas de las publicaciones. Estos datos permitirán enriquecer aún más la evaluación de la producción científica, al ofrecer una visión más detallada de las áreas de investigación en las que la institución está involucrada. La incorporación de estos elementos facilitará la identificación de tendencias emergentes y colaboraciones interdisciplinarias, lo que ampliará las posibilidades de análisis y permitirá realizar estudios más completos sobre el impacto de la investigación.

Además, el uso del modelado *Data Vault* proporciona la flexibilidad necesaria para integrar estas nuevas entidades y sus relaciones sin afectar la estructura existente ni comprometer la calidad de los datos ya procesados. Esto permite ampliar el alcance del análisis incorporando información de fuentes adicionales y adaptándose a la evolución de la producción científica institucional.

El enfoque adoptado para la integración de datos de OpenAIRE y OpenAlex puede ser un factor clave en el desarrollo de repositorios digitales de próxima generación. Tal como se menciona en el informe de COAR (Bollini et al., 2017), los repositorios deben evolucionar y adaptarse a nuevas tecnologías y estándares que mejoren el acceso y el intercambio de la investigación. La integración de datos de diferentes fuentes, como las métricas de

citas de OpenAlex y las métricas de visibilidad como vistas y descargas de OpenAIRE, puede ser un paso fundamental en este proceso.

Al ofrecer una visión más completa del impacto de la producción científica, este enfoque permite a los repositorios presentar métricas relevantes obtenidas de múltiples fuentes, enriqueciendo la información disponible para los usuarios. Esto no solo facilita el acceso y la difusión de la investigación, sino que también contribuye a la mejora de los servicios ofrecidos por los repositorios, alineándolos con la visión de transformar estos sistemas en componentes clave del ecosistema académico global.

Además, la integración de datos de diferentes fuentes y el análisis de métricas puede promover la innovación al proporcionar herramientas más efectivas para medir y gestionar el impacto de la investigación de manera más precisa. Así, se está contribuyendo al objetivo más amplio de crear un sistema de comunicación académica más abierto y accesible, que reduzca las barreras económicas y fomente la equidad en el acceso al conocimiento, como se plantea en las recomendaciones del Grupo de Trabajo sobre Repositorios de Nueva Generación.

Asimismo, resulta pertinente mencionar herramientas complementarias como HERA (Herramienta de Evaluación de Revistas Académicas), que se centra en asistir al proceso de valoración de revistas y artículos a partir del ingreso de un identificador como el DOI (Porto, 2021). A diferencia del enfoque propuesto en este trabajo, que parte de una recuperación masiva basada en la afiliación institucional de los autores para construir tableros analíticos, HERA actúa como un servicio de consulta puntual, recuperando y agregando métricas desde múltiples fuentes para un recurso específico (Carletti & Villarreal, 2024). Esta diferencia metodológica no solo resalta la complementariedad entre ambas propuestas, sino que también abre la posibilidad de incorporar nuevas herramientas como fuentes adicionales en futuros desarrollos, enriqueciendo la evaluación institucional con perspectivas innovadoras a partir de consultas focalizadas sobre recursos de interés.

## 6 Conclusiones

La estrategia de integración de datos académicos institucionales propuesta en este artículo ha demostrado ser eficaz para obtener una visión más completa y unificada del impacto de la producción científica institucional. Mediante la combinación de datos provenientes de diversas fuentes como OpenAIRE, OpenAlex y COAR, hemos logrado consolidar métricas clave — incluyendo citas, vistas y descargas — en una tabla unificada de publicaciones académicas. Esta unificación no solo facilita la toma de decisiones informadas sobre la evaluación del rendimiento institucional, sino que también promueve la implementación de métricas responsables.

Uno de los principales aportes de este trabajo radica en la metodología diseñada para abordar la multiplicidad y heterogeneidad de las fuentes de datos. El sistema de integración, basado en herramientas de código abierto como Kedro y dbt, y un modelo de datos híbrido que combina *Data Warehouse* (Kimball & Ross) con *Data Vault 2.0* (Linstedt & Olschimke), ha permitido gestionar la complejidad de los metadatos y asegurar la trazabilidad y consistencia de la información. La adopción del enfoque ELT (*Extract, Load, Transform*) ha resultado particularmente beneficiosa para el manejo de grandes volúmenes de datos académicos dinámicos, garantizando su conservación original y permitiendo análisis retrospectivos y una mayor trazabilidad de todas las transformaciones realizadas.

A partir de los resultados alcanzados, se ha evidenciado cómo la integración permite discernir con mayor precisión la producción científica institucional relevante, excluyendo elementos que podrían distorsionar el análisis del impacto contemporáneo. La implementación de un tablero de control interactivo en Apache Superset, alimentado por la tabla consolidada en la capa de *Data Mart*, proporciona una herramienta robusta para explorar y visualizar la producción científica, comparar indicadores entre fuentes y analizar su evolución temporal. Este estudio confirma la aplicabilidad y flexibilidad del modelo propuesto para adaptarse a diferentes entornos institucionales y sistemas de información. Los desafíos inherentes a la integración de datos heterogéneos, como la desambiguación de entidades y la diferencia en la calidad de los metadatos, fueron abordados mediante la normalización precisa de identificadores persistentes como los DOI y el uso de vocabularios controlados como el RTV de COAR.

Este enfoque no solo se alinea con recomendaciones de la bibliometría crítica y las métricas responsables, sino que también demuestra una aplicabilidad concreta para instituciones que deseen evaluar su producción científica de forma integral. En el futuro, la incorporación de tópicos temáticos, fuentes adicionales y modelos predictivos

ampliará aún más el alcance del análisis. En suma, el modelo presentado ofrece una base sólida y adaptable para avanzar hacia sistemas de información académica más robustos, abiertos y estratégicamente orientados.

## Referencias


Add seeds to your DAG. (2025, Abril 3). dbt Developer Hub. Recuperado el Abril 4, 2025, de https://docs.getdbt.com/docs/build/seeds

Aghassibake, N., Castello, O. G., Gujilde, P., & Rabun, S. (2023). Visualizing institutional activity using persistent identifier metadata. *Information Services & Use, 43*(3-4), 335–342. https://doi.org/10.3233/ISU-230218

Albuquerque, P. C. (2024a). *PabloDeAlbu/dbt-scholar* [Software]. GitHub. https://github.com/PabloDeAlbu/dbt-scholar

Albuquerque, P. C. (2024b). *PabloDeAlbu/kedro-scholar* [Cuaderno Jupyter]. GitHub. https://github.com/PabloDeAlbu/kedro-scholar

Albuquerque, P. C., Villarreal, G. L., & De Giusti, M. R. (2021, Junio 22–25). *Proposal of a data warehouse for scholarly institutions built on institutional repositories* [Objeto de conferencia]. IX Jornadas de Cloud Computing, Big Data & Emerging Topics, La Plata, Buenos Aires, Argentina. http://sedici.unlp.edu.ar/handle/10915/125161

Albuquerque, P. C., Villarreal, G. L., & De Giusti, M. R. (2022, Octubre 3-7). *WebID como base para el desarrollo de una marca personal en repositorios institucionales* [Objeto de conferencia]. XI Conferencia Internacional de Bibliotecas y Repositorios Digitales (BIREDIAL-ISTEC), Costa Rica. http://sedici.unlp.edu.ar/handle/10915/145739

Albuquerque, P. C., Villarreal, G. L., & De Giusti, M. R. (2023, Octubre 18-20). *Modelo dimensional para la medición de la producción académica* [Objeto de conferencia]. XII Conferencia Internacional de Bibliotecas y Repositorios Digitales (BIREDIAL-ISTEC), Montevideo, Uruguay. http://sedici.unlp.edu.ar/handle/10915/161906

Apache Superset. (2025). Apache Superset™ is an open-source modern data exploration and visualization platform. Recuperado el Abril 4, 2025, de https://superset.apache.org/

Bollini, A., Knoth, P., Perakakis, P., Rodrigues, E., Shearer, K., Sompel, V. de, & Walk, P. (2017). *Next generation repositories: Behaviours and technical recommendations of the COAR Next Generation Repositories Working Group* (Version 2) [Original report]. Zenodo. https://doi.org/10.5281/zenodo.8077381

Cabezas-Clavijo, A., & Torres-Salinas, D. (2021). Bibliometric reports for institutions: Best practices in a responsible metrics scenario. *Frontiers in Research Metrics and Analytics, 6*, Article e696470. https://doi.org/10.3389/frma.2021.696470

Carletti, E., Rucci, E., & Villarreal, G. L. (2024, Octubre 22-24). *HERA 2.0: Más funcionalidad para la evaluación de recursos académicos* [Objeto de conferencia]. XIII Conferencia Internacional de Bibliotecas y Repositorios Digitales (BIREDIAL-ISTEC), Santiago de Chile, Chile. http://sedici.unlp.edu.ar/handle/10915/177287

Ciuciu-Kiss, J. T., & Garijo, D. (2024, May 27). Assessing the overlap of science knowledge graphs: A quantitative analysis [Conference paper]. International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs, Hersonissos, Crete, Greece. In G. Rehm, S. Dietze, S. Schimmler, & F. Krüger (Eds.), *Natural scientific language processing and research knowledge graphs, Lecture Notes in Computer Science* (Vol. 14770, pp. 171-185). Springer. https://doi.org/10.1007/978-3-031-65794-8_11

Cuartas, G. V., Tirado, A. U., Restrepo-Quintero, D., Gutiérrez, J. O., Pallares, C., Gómez-Molina, H. F., Suárez-Tamayo, M., & Calle, J. (2019). Hacia un modelo de medición de la ciencia desde el Sur Global: Métricas responsables. *Palabra Clave, 8*(2), Artículo e068. https://doi.org/10.24215/18539912e068

Data catalog. (2025). Kedro. Recuperado el Julio 22, 2025, de https://docs.kedro.org/en/1.0.0/catalog-data/introduction/

Dhaouadi, A., Bousselmi, K., Gammoudi, M. M., Monnet, S., & Hammoudi, S. (2022). Data warehousing process modeling from classical approaches to new trends: Main features and comparisons. *Data, 7*(8), Article 113. https://doi.org/10.3390/data7080113

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021, September). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Filtering search results. (2025). OpenAIRE Graph Documentation. Recuperado el Julio 22, 2025, de https://graph.openaire.eu/docs/10.3.0/apis/graph-api/searching-entities/filtering-search-results/

Harder, R. (2024, June). Using Scopus and OpenAlex APIs to retrieve bibliographic data for evidence synthesis: A procedure based on Bash and SQL. *MethodsX, 12*, Article 102601. https://doi.org/10.1016/j.mex.2024.102601

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. de, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys, 54*(4), Article 71. https://doi.org/10.1145/3447772

Kimball, R., & Ross, M. (2013). *The data warehouse lifecycle toolkit* (3rd ed.). John Wiley & Sons.

Linstedt, D., & Olschimke, M. (2015). *Building a scalable data warehouse with Data Vault 2.0* (1st ed.). Morgan Kaufmann.

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., & Principe, P. (2019). *The OpenAIRE research graph data model* (Version 1.3) [Original report]. Zenodo. https://doi.org/10.5281/zenodo.2643199

Öztürk, O., Kocaman, R., & Kanbach, D. K. (2024). How to design bibliometric research: An overview and a framework proposal. *Review of Managerial Science, 18*, 3333-3361. https://doi.org/10.1007/s11846-024-00738-0

Priem, J., Piwowar, H., & Orr, R. (2022, May 4). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* [Preprint arXiv]. Submitted to the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada, Spain. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Searching entities. (2025). OpenAIRE Graph Documentation. Recuperado el Julio 22, 2025, de https://graph.openaire.eu/docs/apis/graph-api/searching-entities/

Silva, V. S., Matas, L., Moreira, T., & Segundo, W. C. (2022). An ETL strategy for integrating the LA Referencia platform and VIVO for the Brazilian CRIS. *Procedia Computer Science, 211,* 111-117. https://doi.org/10.1016/j.procs.2022.10.182

Tomczyńska, A., Ostrowska, S., Protasiewicz, J., & Podwysocki, E. (2023, June 15). *Beyond CRIS: A research and higher education information system in Poland* [Paper]. EUNIS 2023 Annual Conference, Vigo, Spain. http://hdl.handle.net/11366/2477

Universidad Nacional de La Plata. (2025). OpenAlex. Recuperado el Abril 4, 2025, de https://openalex.org/institutions/i874386039

Use a Jupyter notebook for Kedro project experiments. (2024). Kedro. Recuperado el Abril 4, 2025, de https://docs.kedro.org/en/stable/notebooks_and_ipython/kedro_and_notebooks.html

Works overview: Schema reference for Works entities. (2025). OpenAlex. Recuperado el Abril 4, 2025, de https://docs.openalex.org/api-entities/works/work-object

## Datos de publicación

**Pablo César de Albuquerque**
Licenciado en Sistemas

Universidad Nacional de La Plata, La Plata, BA, Argentina
Comisión de Investigaciones Científicas, La Plata, BA, Argentina

pablo@sedici.unlp.edu.ar

https://orcid.org/0000-0001-5277-1665

Es Licenciado en Sistemas por la Universidad Nacional de La Plata (UNLP) y actualmente cursa el Doctorado en Ciencias Informáticas en la Facultad de Informática de la misma universidad. Desarrolla su trabajo de investigación en PREBI-SEDICI (UNLP) y en el Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC).

Su tesis doctoral se centra en el diseño e implementación de un data warehouse académico que integre múltiples fuentes para medir la visibilidad e impacto de la producción científica institucional. Sus áreas de interés incluyen la ciencia de datos, bibliometría, repositorios digitales y la gestión de información académica.

**Gonzalo Luján Villarreal**
Doctor en Ciencias Informáticas

Universidad Nacional de La Plata, La Plata, BA, Argentina
Comisión de Investigaciones Científicas, La Plata, BA, Argentina

gonzalo@sedici.unlp.edu.ar

https://orcid.org/0000-0002-3602-8211

Es Doctor en Ciencias Informáticas por la Universidad Nacional de La Plata (UNLP). Actualmente se desempeña como Director de PREBI-SEDICI UNLP y Director del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas (CIC) de la Provincia de Buenos Aires.

En el ámbito académico, es docente en la Facultad de Informática de la UNLP, donde dicta cursos de programación, programación orientada a objetos y programación concurrente, así como asignaturas de posgrado relacionadas con métricas científicas y repositorios digitales.

Además, es Coordinador Técnico de revistas científicas de la UNLP y responsable de la gestión de los portales de revistas, congresos, libros y del Repositorio de Datos de Investigación de la universidad. Sus intereses de investigación incluyen bibliotecas digitales, repositorios, desarrollo y ingeniería de software, y simulación de eventos discretos.

**Dirección de correspondencia del autor principal**
131, 723, 1900, La Plata, Argentina.

**Originalidad**
Declaro que el texto es original y no está siendo evaluado en ninguna otra publicación.

**Preprints**
El manuscrito no fue enviado a ninguna plataforma de preprints.

**Información sobre el trabajo**
Este manuscrito constituye un avance de la tesis doctoral en curso dentro del Doctorado en Ciencias Informáticas de la Facultad de Informática de la Universidad Nacional de La Plata (UNLP). La investigación se centra en el diseño e implementación de un data warehouse académico para la

integración de métricas de impacto institucional. En caso de que el artículo sea publicado antes de la defensa de la tesis, el mismo será debidamente citado en el documento final de la misma.

El manuscrito es financiado a través de la beca de Postgrado “BDOC19 CP/TP”, otorgada por la COMISION DE INVESTIGACIONES CIENTIFICAS DE LA PROVINCIA DE BUENOS AIRES (CIC).

## Agradecimientos

Los autores expresan su agradecimiento al Centro de Servicios en Gestión de Información (CESGI - CIC) y a la Dirección PREBI-SEDICI de la Universidad Nacional de La Plata (UNLP) por el apoyo institucional, el acceso a las infraestructuras de datos y el entorno de investigación brindado para el desarrollo de este trabajo.

## Contribución de los autores

Concepción y preparación del manuscrito: P. C. de Albuquerque, G. L. Villarreal.

Recogida de datos: P. C. de Albuquerque.

Análisis de datos: P. C. de Albuquerque.

Discusión de los resultados: P. C. de Albuquerque, G. L. Villarreal.

Revisión y aprobación: P. C. de Albuquerque, G. L. Villarreal.

## Uso de inteligencia artificial

Para optimizar la legibilidad y la estructura narrativa del trabajo, se utilizó la asistencia de la inteligencia artificial Gemini 1.5 Flash. El proceso consistió en proporcionar a la herramienta fragmentos de texto extensos junto con contexto específico sobre la disciplina (Ciencias de la Información y Bibliometría). Se utilizaron prompts tales como: 'Ayúdame a desarrollar este bloque de texto en dos o tres párrafos más breves para favorecer la narrativa y la claridad, considerando que el tema es la integración de métricas mediante Data Warehouse'. Tras las sugerencias de segmentación y redacción propuestas por la IA, los autores realizaron una revisión crítica para asegurar la precisión técnica y el rigor científico del contenido final.

## Financiación

El manuscrito es financiado a través de la beca de Postgrado “BDOC19 CP/TP”, otorgada por la COMISIÓN DE INVESTIGACIONES CIENTÍFICAS DE LA PROVINCIA DE BUENOS AIRES (CIC).

## Autorización para utilizar imágenes

No aplicable.

## Aprobación del comité de ética de la investigación

No aplicable.

## Conflicto de intereses

No aplicable.

## Declaración de Disponibilidad de Datos

Los datos y el código fuente que respaldan este estudio se encuentran disponibles en repositorios de acceso abierto:

Albuquerque, P. C. (2024a). PabloDeAlbu/dbt-scholar [Software]. https://github.com/PabloDeAlbu/dbt-scholar

Albuquerque, P. C. (2024b). PabloDeAlbu/kedro-scholar [Software]. https://github.com/PabloDeAlbu/kedro-scholar

## Licencia de uso

**Editor**

Publicada por el University Library System de la Universidad de Pittsburgh. Responsabilidad compartida con las universidades asociadas. Las ideas expresadas en este artículo son las de los autores y no representan necesariamente las opiniones de los editores o de la universidad.

**Editores**

Lúcia da Silveira, Fabiano Couto Corrêa da Silva y Laura Vilela Rodrigues Rezende.

**Histórico**

Recibido: 14-04-2025 – Aprobado: 09-03-2026 – Publicado: 15-05-2026.