UN FRAMEWORK FLEXIBLE PARA LA MEJORA DE METADATOS EN REPOSITORIOS INSTITUCIONALES CON DATOS DE OPENAIRE Y OPENALEX

Pablo César de Albuquerque
Universidad Nacional de La Plata (UNLP); Comisión de Investigaciones Científicas (CIC), Argentina
pablo@sedici.unlp.edu.ar
https://orcid.org/0000-0001-5277-1665

Gonzalo Luján Villarreal
Universidad Nacional de La Plata (UNLP); Comisión de Investigaciones Científicas (CIC), Argentina
gonzalo@sedici.unlp.edu.ar
https://orcid.org/0000-0002-3602-8211

DOI: *10.22477/xiv.biredial.406*
EJE TEMÁICO: *Infraestructura tecnológica*

RESUMEN

Este trabajo presenta una estrategia para enriquecer y analizar los metadatos de un repositorio institucional mediante la integración de fuentes abiertas como OpenAIRE y OpenAlex. Partiendo de la premisa de que estas fuentes ofrecen perspectivas complementarias sobre la producción científica, se propone su combinación para construir una visión más integrada y enriquecida. El enfoque se basa en un modelado flexible con Data Vault, que permite una integración escalable de datos, normalizando identificadores y vinculando entidades clave. El proceso se organiza en tres etapas: recolección, normalización e integración. Durante la recolección, se aplican filtros que permiten obtener datos pertinentes desde ambas fuentes, considerando tanto la afiliación institucional como el repositorio local. La integración cruza información, combinando métricas de impacto y visibilidad. Se presentan resultados obtenidos para el repositorio SEDICI de la UNLP, incluyendo publicaciones enriquecidas, autores institucionales identificados y publicaciones candidatas a ser incorporadas desde OpenAlex. Finalmente, se discute el potencial de extender esta estrategia a otras fuentes, incluyendo sistemas internos, para mejorar la cobertura y calidad de los metadatos. La propuesta promueve un uso estratégico del repositorio como herramienta de gestión, evaluación y difusión de la producción científica institucional.

Palabras-clave: Bibliometría, Repositorios institucionales, Data Warehouse, Identificadores persistentes.

ABSTRACT

(ver resumo em inglês com autores)

Keywords: Bibliometrics, Institutional repositories, Data Warehouse, Persistent identifiers.

INTRODUCCIÓN

La creciente disponibilidad de fuentes abiertas que ofrecen datos bibliográficos, métricas de impacto y visibilidad plantea una oportunidad estratégica para los repositorios institucionales (IR): mejorar la calidad de sus metadatos, ampliar la cobertura de sus colecciones y ofrecer servicios de valor agregado. Tal como señala el informe de COAR (Bollini *et al.*, 2017), los repositorios deben evolucionar y adaptarse a nuevas tecnologías y estándares que favorezcan el acceso abierto, la interoperabilidad y el intercambio efectivo de información científica.

En este escenario, la bibliometría se presenta como una herramienta clave para analizar y comprender la producción científica de una institución. A través de indicadores cuantitativos como el número de publicaciones, citas recibidas, y la identificación de autores e instituciones más prolíficas, esta disciplina permite medir el impacto académico y detectar tendencias emergentes en distintas áreas del conocimiento (Öztürk *et al.*, 2024; Donthu *et al.*, 2021). Su aplicación, combinada con datos integrados desde múltiples fuentes, puede aportar insumos valiosos para la toma de decisiones estratégicas en políticas de investigación y desarrollo.

Justamente, la integración de datos provenientes de fuentes como OpenAlex —que ofrece información bibliográfica y métricas de citas— y OpenAIRE —que proporciona indicadores de uso como vistas y descargas— permite construir una visión más completa del impacto y la visibilidad de la producción científica institucional. Esta información no solo enriquece los registros disponibles para los usuarios finales, sino que también aporta valor a los equipos de gestión académica, permitiendo evaluar y comunicar con mayor precisión el alcance de sus publicaciones.

Este enfoque responde a una visión más amplia: posicionar a los repositorios como infraestructuras activas dentro del ecosistema de ciencia abierta. La capacidad de recolectar, normalizar y presentar datos integrados habilita nuevos servicios, promueve una mayor equidad en el acceso al conocimiento, y contribuye a la transición hacia un sistema de comunicación académica más transparente, inclusivo y eficiente, tal como se plantea en las recomendaciones del Grupo de Trabajo sobre Repositorios de Nueva Generación.

OBJETIVO DEL TRABAJO

Este artículo propone un enfoque sistemático para la recolección e integración de datos provenientes de fuentes abiertas y confiables, como OpenAIRE y OpenAlex, con el fin de enriquecer los metadatos existentes en repositorios institucionales e incorporar nuevos registros cuando sea necesario. La metodología presentada contempla mecanismos de conciliación entre la información externa y los registros locales, priorizando el uso de identificadores persistentes, como ROR, DOI y ORCID, para garantizar la interoperabilidad, la trazabilidad y la posibilidad de replicación en otras instituciones.

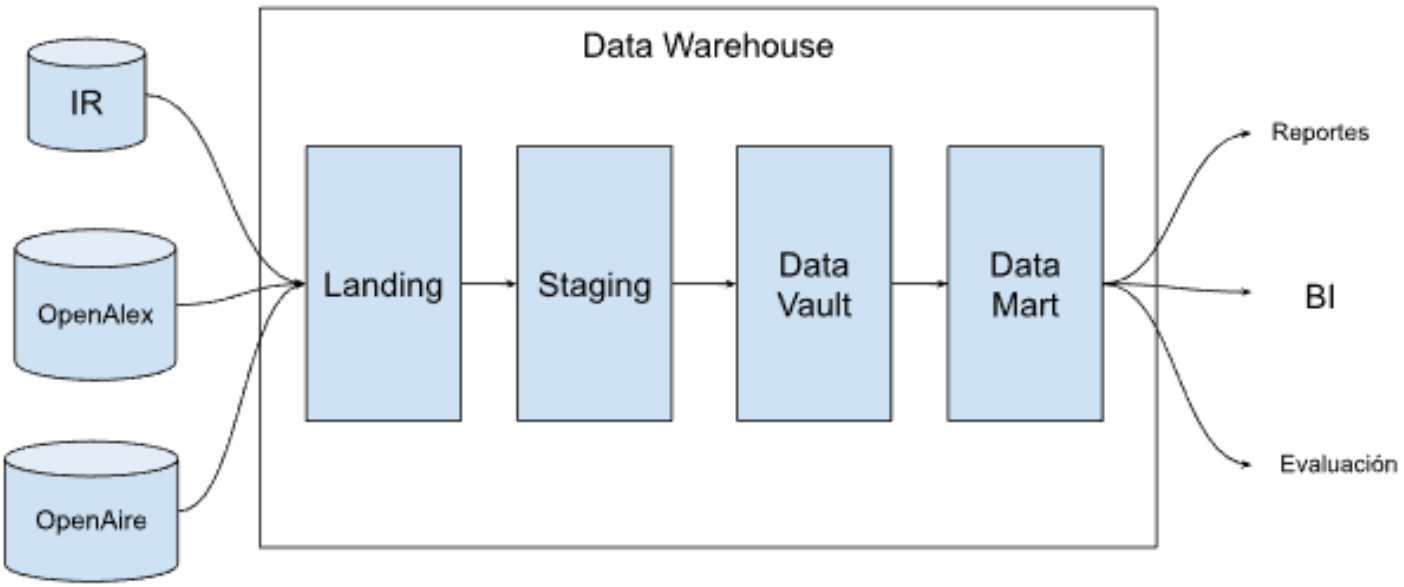
Para alcanzar este objetivo, se adopta una arquitectura basada en un Data Warehouse modelado con enfoque Data Vault, que permite organizar y escalar la integración de múltiples fuentes de datos. La implementación se apoya en herramientas de orquestación y transformación como Kedro y dbt, que facilitan la trazabilidad y el versionado de los procesos. Como caso de aplicación, se presenta una experiencia desarrollada sobre el repositorio institucional de la Universidad Nacional de La Plata (SEDICI), mostrando cómo este enfoque puede contribuir a mejorar la calidad de los metadatos y a ofrecer indicadores más precisos sobre la visibilidad e impacto de la producción académica.

METODOLOGÍA

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 los datos, desde la extracción hasta la presentación final. La Imagen 1 ilustra esta arquitectura general.

Imagem 1 - Diagrama general de la arquitectura del sistema, con sus cuatro capas principales: *landing*, *staging*, *data vault* y *data mart*



Fuente: Elaboración propia.

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.

STAGING

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

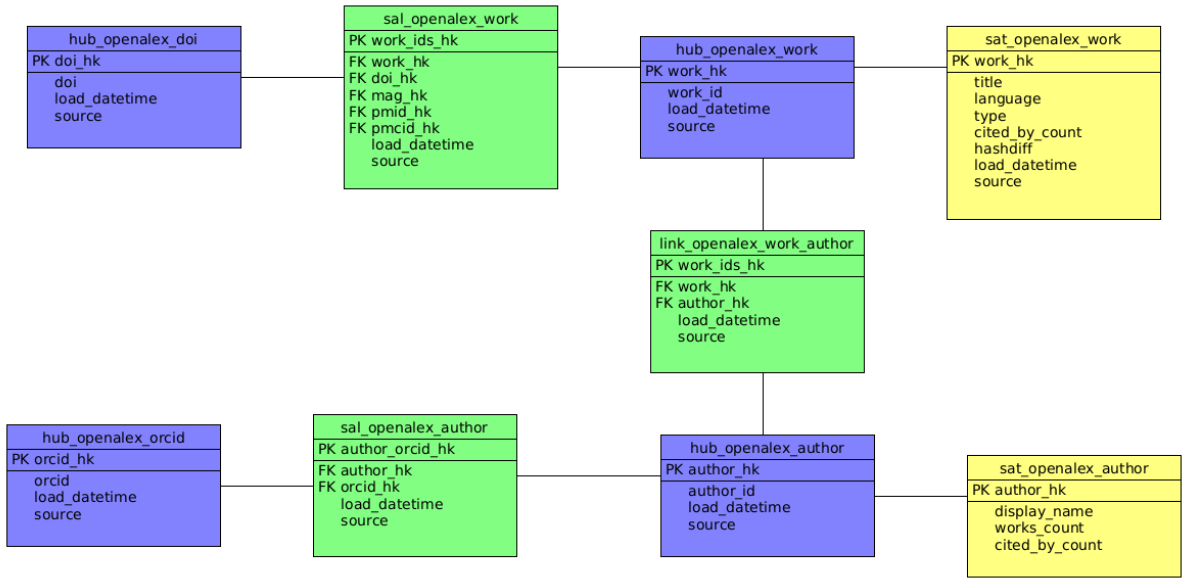
En esta etapa también se generan claves primarias para identificar de manera unificada las entidades recuperadas, sin depender de sus claves de negocio (*business keys*). Estas últimas son identificadores relevantes para nuestra investigación, pero pueden variar en su representación según la fuente de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

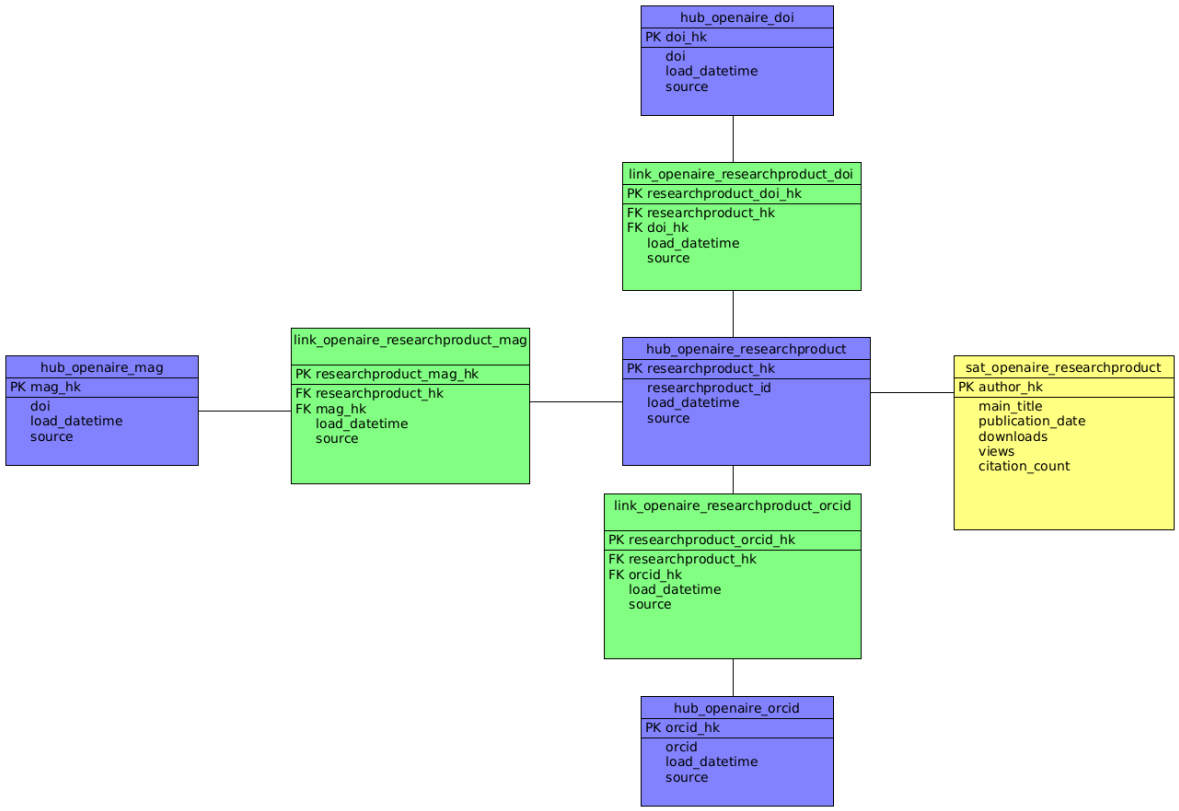
Esta estructura posibilita mantener la integridad histórica de los datos y facilita la actualización ante cambios en las fuentes o en los requisitos del análisis. La Imagen 2 muestra un modelo simplificado aplicado a datos de OpenAlex, mientras que la Imagen 3 presenta el correspondiente a OpenAIRE.

Imagen 2 - Modelo simplificado de Data Vault utilizado para los datos obtenidos de OpenAlex



Fuente: Elaboración propia.

Imagen 3 - Modelo simplificado de Data Vault utilizado para los datos obtenidos de OpenAire.



Fuente: Elaboración propia.

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:

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

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.

CONSTRUCCIÓN DEL MODELO DIMENSIONAL

El *data mart* se construye a partir de los datos previamente integrados en la capa de data vault, utilizando un enfoque dimensional. Esto implica organizar la información en una tabla de hechos central, acompañada por distintas dimensiones que contextualizan los datos y permiten su análisis desde múltiples perspectivas. El objetivo principal es consolidar métricas clave, como cantidad de citas, vistas y descargas, en una estructura accesible para los usuarios finales.

La consolidación de estos datos se logra mediante la intersección de registros provenientes de fuentes abiertas como OpenAlex y OpenAIRE, utilizando identificadores persisten-

tes, como el DOI, para vincular publicaciones equivalentes. Este proceso permite unificar los distintos indicadores aportados por cada fuente y agruparlos en un único registro por publicación. En otras palabras, se parte de publicaciones extraídas por separado, se normalizan sus identificadores, y luego se cruzan para formar una única tabla de hechos que combina las métricas relevantes de cada origen.

Este modelo dimensional se apoya en un enfoque híbrido que combina los principios de Data Vault 2.0 con la metodología propuesta por Kimball & Ross (2013) para el desarrollo de Data Warehouses. El uso de hechos y dimensiones, heredado del enfoque dimensional clásico, facilita el análisis transversal y comparativo a lo largo del tiempo, mientras que la base en Data Vault asegura trazabilidad, flexibilidad y consistencia en el tratamiento de datos heterogéneos.

El diseño del sistema y su implementación con herramientas de código abierto permite una adaptación sencilla a otras instituciones. Al estar basado en identificadores persistentes como DOI, ORCID y ROR, el modelo puede generalizarse más allá del caso específico estudiado.

La literatura reciente ha destacado la utilidad de estos enfoques en el ámbito académico. Diversos trabajos han explorado la construcción de Data Warehouses a partir de datos de repositorios institucionales como una solución efectiva para unificar la información disponible y mejorar la toma de decisiones (de Albuquerque *et al.*, 2021; de Albuquerque *et al.*, 2023). Estos antecedentes validan la propuesta de este artículo y refuerzan la aplicabilidad del modelo, subrayando su capacidad de adaptación a diferentes contextos institucionales y sistemas de información.

FUENTES DE DATOS

Este trabajo se basa en la integración de dos fuentes abiertas de alta calidad que actúan como science knowledge graphs: OpenAIRE Graph (Manghi *et al.*, 2019) y OpenAlex (Priem *et al.*, 2022). Estas plataformas permiten representar entidades científicas y sus relaciones, como obras, autores, instituciones o fuentes, mediante identificadores persistentes como DOI, ORCID o ROR, y ofrecen acceso a través de APIs diseñadas para su procesamiento automatizado (Hogan *et al.*, 2021; Ciuciu-Kiss & Garijo, 2024).

OpenAIRE Graph consolida información proveniente de repositorios institucionales, sistemas CRIS y editores académicos, siendo particularmente útil para recuperar datos depositados en el ecosistema europeo de acceso abierto. Por su parte, OpenAlex, sucesor del Microsoft Academic Graph, organiza su información como un grafo interrelacional y cuenta con una API moderna que facilita la recuperación de datos actualizados y bien estructurados.

La combinación de estas fuentes ofrece una visión más amplia y diversa de la producción científica, pero también plantea desafíos técnicos. Existen diferencias en los esquemas de metadatos, la granularidad de la información, la calidad de los registros, y la evolución de las

estructuras internas. Por ejemplo, OpenAlex ha modificado su modelo de datos con el tiempo: cambió el término venue por source, y actualizó identificadores de entidades, lo que obligó a adaptar el código para mantener la coherencia del sistema (Harder, 2024).

ESTRATEGIA DE RECOLECCIÓN DE DATOS

Extracción desde OpenAIRE Graph

Para recuperar publicaciones asociadas a una institución en OpenAIRE, se utiliza el endpoint */researchProducts*[1], combinando filtros[2] que permiten focalizar los resultados:

- *relOrganizationId*: filtra por organización, utilizando el identificador interno de OpenAIRE.
- *relCollectedFromDatasourceId*: filtra por fuente de datos, como el repositorio institucional.

El parámetro *relOrganizationId* permite recuperar recursos vinculados a una institución específica, aunque OpenAIRE no mantiene identificadores internos para autores ni una relación explícita entre estos y sus afiliaciones. La única forma confiable de validar una afiliación es a través de ORCID, pero esta información no siempre está presente en los registros.

En cambio, el uso de *relCollectedFromDatasourceId* permite enfocar la búsqueda en fuentes particulares asociadas a la institución, como su repositorio. Si bien esto no garantiza que todos los autores estén afiliados, ofrece un indicio fuerte de pertenencia institucional, ya que los repositorios suelen contener producción científica de su comunidad académica.

El identificador de OpenAIRE para una organización se puede obtener a partir de su ROR. Por ejemplo, para la UNLP, cuyo ROR es *https://ror.org/01tjs6929*, se puede consultar su identificador interno con el siguiente endpoint:

https://api.openaire.eu/graph/v1/organizations?pid=https://ror.org/01tjs6929

Luego, para ver qué fuentes de datos están registradas en OpenAIRE para esa organización (como su repositorio institucional), se utiliza:

https://api.openaire.eu/graph/v1/dataSources?relOrganizationId=openorgs____::40b9f835648a3e0d057d6917dd7e54d5

Una vez identificado el repositorio, se puede recuperar su producción científica usando:

https://api.openaire.eu/graph/researchProducts?relCollectedFromDatasourceId=<ID_DEL_REPOSITORIO>

[1] Research products | OpenAIRE Graph Documentation https://graph.openaire.eu/docs/data-model/entities/research-product/
[2] Filtering search results | OpenAIRE Graph Documentation https://graph.openaire.eu/docs/apis/graph-api/searching-entities/filtering-search-results

Este procedimiento permite recolectar datos relevantes vinculados a una institución específica, apoyándose en identificadores persistentes y fuentes registradas.

Extracción desde OpenAlex

OpenAlex, como sucesor del Microsoft Academic Graph, proporciona una API robusta para acceder a una amplia gama de datos científicos, como obras, autores, instituciones y fuentes. La recolección de datos se realiza principalmente a través de su API, utilizando identificadores persistentes como DOI, ORCID y ROR para garantizar la consistencia y la vinculación de las entidades.

El proceso de recolección en OpenAlex comienza con la identificación de los recursos de interés, como publicaciones, autores o instituciones. Por ejemplo, para obtener publicaciones asociadas a una institución, se puede utilizar el endpoint */works*[3], filtrando por el identificador ROR de la organización:

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

Este endpoint permite recuperar todas las obras asociadas a la UNLP, utilizando el ROR como identificador persistente de la institución. La API de OpenAlex también ofrece filtros adicionales, como el tipo de obra, el periodo de publicación, o el número de citas, lo que facilita una recolección más específica según las necesidades del análisis.

Además de los trabajos, OpenAlex proporciona endpoints específicos para acceder a información sobre autores y sus afiliaciones institucionales. Para recolectar datos sobre autores, se puede utilizar el endpoint */authors*[4], filtrando por ORCID o nombre. La API de OpenAlex también permite recuperar información sobre los conceptos (temas de investigación), las revistas y las fuentes, facilitando la integración de un modelo dimensional para análisis más detallados.

ESTRATEGIAS DE INTEGRACIÓN DE FUENTES

Una vez que los datos han sido normalizados y ubicados en la capa de Data Vault, se procede a la integración de las fuentes en la capa de Data Mart. En este proceso, se emplean los DOI como identificadores clave para los recursos institucionales, ya que tanto OpenAIRE como OpenAlex utilizan estos identificadores para vincular publicaciones, aunque con algunas diferencias en su implementación y manejo.

Para integrar los datos recuperados de OpenAIRE y OpenAlex, la unificación se realiza a través de los hubs de DOI en el modelo de Data Vault. Esto permite consolidar las métricas de cada fuente en una única tabla de hechos, facilitando el análisis en el modelo dimensional. Por

[3] Works | OpenAlex technical documentation - https://docs.openalex.org/api-entities/works.
[4] Authors | OpenAlex technical documentation - https://docs.openalex.org/api-entities/authors.

ejemplo, las citas extraídas del satellite de OpenAlex se combinan con las vistas y descargas obtenidas del satellite de OpenAIRE, proporcionando una visión completa de la visibilidad y el impacto de las publicaciones. Este enfoque asegura que los datos de ambas fuentes se estructuren de manera coherente y estén disponibles para análisis posteriores.

Un reto importante en este proceso es cómo manejar las variaciones en los identificadores entre las fuentes. En algunos casos, una misma publicación puede tener múltiples DOI, como en el caso de una versión preprint en arXiv y la versión publicada en una revista. OpenAlex, sin embargo, solo almacena un único DOI[5], priorizando el correspondiente a la versión final. Esto puede llevar a la pérdida de información si no se realiza un proceso adecuado de reconciliación.

Por otro lado, OpenAIRE aplica un proceso de deduplicación que permite agrupar distintos identificadores persistentes (PID) para un mismo recurso, lo que facilita la consolidación de datos provenientes de diferentes fuentes y evita la duplicación de registros. Este proceso es crucial para garantizar que los datos integrados sean precisos y reflejen la realidad de la producción científica.

Una vez que los datos han sido procesados y las discrepancias de identificadores han sido resueltas, la integración final se realiza en la capa de Data Mart, donde los datos se estructuran en modelos dimensionales, listos para su consulta y análisis. El modelo de Data Mart permite que los usuarios accedan fácilmente a métricas clave de las publicaciones, tales como citas, vistas y descargas, y que estos datos sean utilizados para el análisis de la producción científica de manera efectiva.

RESULTADOS

A partir de los datos recuperados desde OpenAIRE, se identificaron 159.732 publicaciones pertenecientes al repositorio institucional, que representan el 96% del total de registros disponibles para esta prueba (166.350). Estos registros fueron enriquecidos con metadatos adicionales provistos por OpenAIRE, incluyendo vistas, descargas y otras métricas relevantes, lo que mejora significativamente la calidad y profundidad de la información disponible en el repositorio.

El análisis de OpenAlex permitió identificar 19.043 autores vinculados institucionalmente, filtrados por su afiliación a la UNLP mediante identificadores ROR. A partir de su producción científica, se localizaron 24.723 publicaciones con DOI válidos que no están actualmente depositadas en el repositorio institucional. Estas publicaciones se presentan como candidatas a ser incorporadas, permitiendo ampliar la cobertura del repositorio con producción científica relevante que actualmente no se encuentra registrada.

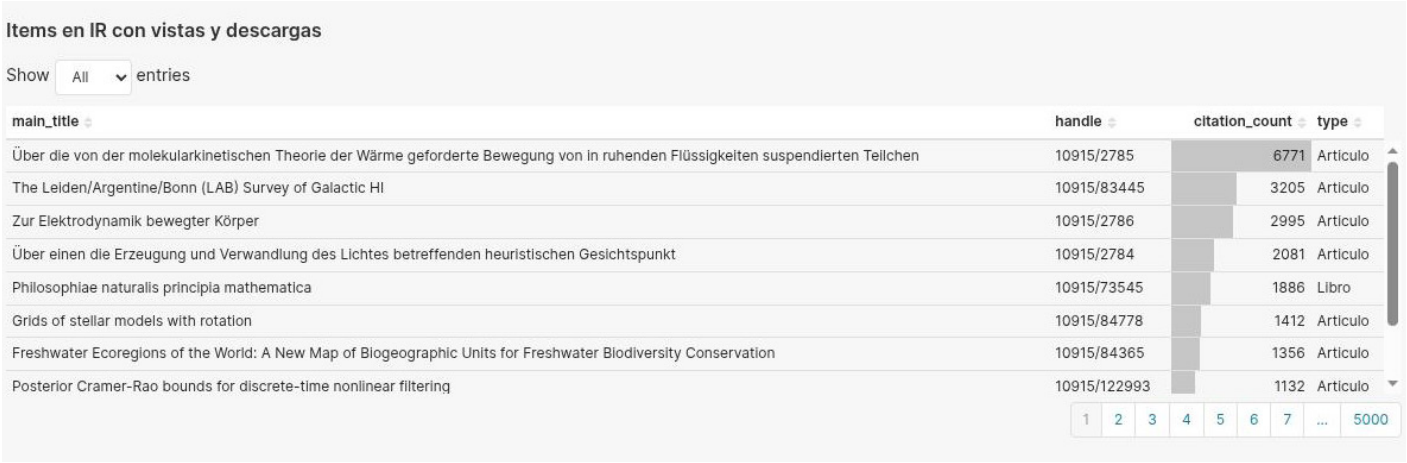
Para analizar el impacto y la cobertura de la producción institucional, se construyeron

[5] Work object | OpenAlex technical documentation https://docs.openalex.org/api-entities/works/work-object#doi

visualizaciones a partir de los datos integrados en el modelo dimensional. A continuación, se presentan dos vistas generadas con Apache Superset[6] que permiten explorar distintos aspectos de los resultados.

La Imagen 4 muestra una tabla con las publicaciones del repositorio institucional (IR) que han sido enriquecidas con citas provenientes de OpenAIRE. Los ítems están ordenados de forma descendente según la cantidad de citas, lo que permite identificar rápidamente cuáles son los trabajos más referenciados dentro del conjunto analizado. Esta visualización no solo aporta información sobre el impacto relativo de las publicaciones, sino que también facilita el acceso a cada ítem a través de su identificador handle.

Imagen 4 - Publicaciones en el Repositorio Institucional enriquecidas con OpenAIRE.



Items en IR con vistas y descargas

Show All entries

main_title	handle	citation_count	type
Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen	10915/2785	6771	Articulo
The Leiden/Argentine/Bonn (LAB) Survey of Galactic HI	10915/83445	3205	Articulo
Zur Elektrodynamik bewegter Körper	10915/2786	2995	Articulo
Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt	10915/2784	2081	Articulo
Philosophiae naturalis principia mathematica	10915/73545	1886	Libro
Grids of stellar models with rotation	10915/84778	1412	Articulo
Freshwater Ecoregions of the World: A New Map of Biogeographic Units for Freshwater Biodiversity Conservation	10915/84365	1356	Articulo
Posterior Cramer-Rao bounds for discrete-time nonlinear filtering	10915/122993	1132	Articulo

1 2 3 4 5 6 7 ... 5000

Fuente: Elaboración propia

La Imagen 5 presenta un listado de publicaciones candidatas a ser incorporadas al repositorio. Estos trabajos fueron recuperados desde OpenAlex y corresponden a autores institucionales cuya producción aún no se encuentra en el IR. Además del título y el número de citas, se incluye un campo que indica si el recurso se encuentra disponible en algún repositorio con acceso al texto completo. Esta tabla constituye una herramienta clave para la curaduría proactiva del repositorio, ya que permite priorizar obras relevantes que actualmente no están depositadas.

[6] Superset | https://superset.apache.org/.

Imagen 5 - Autores institucionales con publicaciones candidatas a ser incorporadas desde OpenAlex.

Autores institucionales en OpenAlex

Show All entries

author_id	display_name_y	works_count	cited_by_count
https://openalex.org/A5046725106	C. Padilla Aranda	2244	123934
https://openalex.org/A5101793588	F. Alonso	1543	82057
https://openalex.org/A5063988052	C. Alpigiani	1425	62036
https://openalex.org/A5113541511	Roland Boese	1386	24498
https://openalex.org/A5019936936	Norma E. Sánchez	1257	431
https://openalex.org/A5067822505	José M. Lorenzo	1167	44320
https://openalex.org/A5051586154	A. Alonso	1166	93221
https://openalex.org/A5026556975	P. Buchholz	1149	75274

1 2 3 4 5 6 7 ... 1905

Fuente: Elaboración propia

El código desarrollado para este trabajo se encuentra disponible en dos repositorios públicos (de Albuquerque, 2025a, 2025b). En ellos se implementa el proceso completo de recolección, integración y análisis de los datos. Kedro se utilizó para estructurar el flujo de trabajo en pipelines reproducibles, mientras que dbt permitió definir y documentar las transformaciones necesarias para organizar los datos en un modelo analítico. Ambas herramientas, de código abierto, facilitaron una implementación modular y mantenible del framework propuesto.

CONCLUSIONES

Los resultados obtenidos permiten evaluar el valor complementario de OpenAIRE y OpenAlex en el proceso de enriquecimiento de metadatos institucionales. Mientras que OpenAIRE brinda acceso a recursos depositados en repositorios institucionales, incluyendo trabajos de digitalización y preservación, OpenAlex permite una identificación más precisa de la producción vinculada a autores institucionales activos.

Esta diferencia sugiere que, para el análisis de producción científica con foco en autores afiliados, OpenAlex resulta más adecuado. Sin embargo, al incluir OpenAIRE se amplía la cobertura y se visibilizan otras dimensiones del trabajo institucional. La combinación de ambas fuentes ofrece así una visión más rica y representativa.

El marco de integración propuesto se destaca por su flexibilidad, permite combinar fuentes con distintos grados de estructuración y cobertura sin comprometer la consistencia del modelo. Esto abre la posibilidad de incorporar nuevas estrategias de recolección, como el acceso directo a la base de datos del repositorio, la exposición vía OAI-PMH, APIs REST o incluso técnicas de scraping. Estas estrategias pueden mejorar la calidad de los datos, especialmente en casos donde OpenAIRE presenta errores de normalización o registros mal fusionados.

La estructura basada en Data Vault facilita la incorporación de nuevas entidades sin afectar los datos existentes. Esto permitirá, en trabajos futuros, incluir elementos como tópicos de investigación (desde OpenAlex) o materias (desde OpenAIRE), para analizar el impacto por áreas temáticas y detectar tendencias emergentes o colaboraciones interdisciplinarias.

Las visualizaciones desarrolladas con Apache Superset muestran cómo esta arquitectura puede ser usada para generar herramientas prácticas para la toma de decisiones. Al permitir ordenar publicaciones por citas, identificar vacíos en el repositorio o priorizar importaciones futuras, se transforma el repositorio en un espacio activo de análisis y curaduría. A su vez, la elección de Superset, software libre, flexible e integrable, refuerza la independencia tecnológica y la sostenibilidad del sistema.

Este enfoque también puede potenciar los procesos de deduplicación e importación masiva que ya existen en SEDICI (Soloaga, 2021), permitiendo incorporar de forma automatizada publicaciones institucionales no depositadas. La información sobre la disponibilidad del texto completo en otros repositorios, provista por OpenAlex, también puede aprovecharse para enriquecer el acceso.

Si bien este trabajo se enfocó en el uso de fuentes abiertas como OpenAIRE y OpenAlex, el enfoque propuesto también permite incorporar datos provenientes de sistemas institucionales que no necesariamente son públicos. Por ejemplo, sistemas de vocabularios controlados y autoridades, como los que gestiona SEDICI en su infraestructura basada en Drupal, pueden aportar metadatos normalizados sobre personas, temas o unidades académicas. Asimismo, sistemas administrativos internos, como los de gestión de investigación o bases de datos de recursos humanos, pueden proveer información valiosa sobre afiliaciones, cargos o trayectorias. La integración de estas fuentes internas con datos abiertos amplía significativamente las posibilidades analíticas, mejorando la precisión de la representación institucional y permitiendo un análisis más profundo de la producción científica. Este enfoque se alinea con principios de interoperabilidad y estándares como CERIF o los promovidos por COAR, que impulsan la vinculación entre sistemas para una gestión más eficiente y coherente del conocimiento académico. Así, los repositorios institucionales no solo pueden consolidar su rol como depositarios de producción científica, sino también como plataformas activas de articulación entre datos, servicios y políticas institucionales.

Finalmente, al integrar datos de distintas fuentes y presentar métricas de impacto y visibilidad, se contribuye a la construcción de repositorios de nueva generación, alineados con la visión de sistemas abiertos, interoperables y centrados en el acceso equitativo al conocimiento.

BIBLIOGRAFÍA

Bollini, A., Knoth, P., Perakakis, P., Rodrigues, E., Shearer, K., Sompel, V., & Walk, P. (2017, November 28). *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group* (version 2). Zenodo. https://doi.org/10.5281/zenodo.8077381.

Ciuciu-Kiss, J. T., & Garijo, D. (2024). Assessing the Overlap of Science Knowledge Graphs: A Quantitative Analysis. In Goebel, R., Wahlster, W., & Zhou, Z. H. (Eds.). *Natural Scientific Language Processing and Research Knowledge Graphs* (pp. 171-185). Springer. https://doi.org/10.1007/978-3-031-65794-8_11.

de Albuquerque, P. C. (2025a). *PabloDeAlbu/dbt-cic* [Software]. https://github.com/PabloDeAlbu/dbt-cic.

de Albuquerque, P. C. (2025b). *PabloDeAlbu/kedro-cic* [Software]. https://github.com/PabloDeAlbu/kedro-cic.

de Albuquerque, P. C., Villarreal, G. L., & De Giusti, M. R. (2021, June 22-25). *Proposal of a Data Warehouse for Scholarly Institutions built on Institutional Repositories* [Short papers]. 9th Conference on Cloud Computing Conference, Big Data & Emerging Topics, Modalidad virtual. http://sedici.unlp.edu.ar/handle/10915/125161.

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

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). 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.

Harder, R. (2024). Using Scopus and OpenAlex APIs to retrieve bibliographic data for evidence synthesis. A procedure based on Bash and SQL. *Methods X, 12*, Article 102601, 1-8. https://doi.org/10.1016/j.mex.2024.102601.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G., 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, 71.1-71:37. https://doi.org/10.1145/3447772.

Kimball, R., & Ross, M. (2013). The Data Warehouse Lifecycle Toolkit (3rd ed John Wiley & Sons, Indianapolis).

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., & Principe, P.

(2019, April 17). *The OpenAIRE Research Graph Data Model* (version 1.3). 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). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (No. arXiv:2205.01833). arXiv. https://doi.org/10.48550/arXiv.2205.01833.

Soloaga, I. (2021). Detección de registros académicos duplicados obtenidos desde repositorios digitales [Tesis, Universidad Nacional de La Plata]. Repositorio Institucional de la UNLP. http://sedici.unlp.edu.ar/handle/10915/115229.

ANEXO 1

RESUMEN BIOGRÁFICO DE LOS AUTORES

Pablo César de Albuquerque

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

Gonzalo Luján Villarreal

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

ANEXO 2

REQUERIMIENTOS DE EQUIPO TÉCNICO PARA LA PRESENTACIÓN DE LA PONENCIA

No aplica.