Vol. 6, No. 1, 2026, 1-22. DOI: 10.47909/ijsmc.342

pro-metrics

ARTÍCULO ORIGINAL

BibFusion: paquete en Python para integrar, desduplicar y armonizar registros bibliográficos exportados de Scopus y Web of Science para análisis bibliométrico

BibFusion: A Python package to integrate, deduplicate, and harmonize exported bibliographic records from Scopus and Web of Science for bibliometric analysis

Angelo Britto[1,2], Sebastian Robledo[3,*], Martha Zuluaga[1]

[1] Universidad Nacional de Colombia, Colombia.
[2] MetricSci, Colombia.
[3] Escuela de pregrado, Dirección Académica, Vicerrectoría de Sede, Universidad Nacional de Colombia, Sede la Paz, Cesar, Colombia.
* Autor correspondiente
Email: srobledog@unal.edu.co. **ORCID:** https://orcid.org/0000-0003-4357-4402

RESUMEN

Objetivo. Presentamos BibFusion, un paquete de software en Python que armoniza exportaciones bibliográficas de Scopus y Web of Science en un corpus único, trazable y listo para el análisis, orientado a la investigación bibliométrica y cienciométrica.

Diseño/Metodología/Enfoque. BibFusion ingiere archivos CSV de Scopus y TXT de Web of Science, aplica una normalización sistemática (p. ej., estandarización ASCII y en mayúsculas de títulos y claves SR; análisis de afiliaciones con extracción de país) y, de forma opcional, enriquece los registros mediante resolución basada en DOI contra OpenAlex para recuperar identificadores persistentes (p. ej., IDs de obras en OpenAlex, ORCID cuando está disponible e IDs de autores en OpenAlex). La integración entre bases utiliza una cascada de desduplicación con prioridad al DOI y un respaldo conservador (título-año-primer autor) cuando el DOI falta. Los autores se desambiguan mediante una jerarquía canónica de PersonID (ORCID → OpenAlexAuthorID → nombre normalizado). Las cadenas de citación se depuran y se remapean para preservar vínculos de citación consistentes y la información de revistas/SCImago se consolida mediante reglas basadas en ISSN/EISSN.

Resultados/Discusión. En una demostración sobre una consulta de marketing emprendedor, BibFusion consolida 436 registros de origen en 253 obras principales únicas y materializa un corpus unificado de 8.569 artículos. El conjunto de datos resultante logra una alta completitud de identificadores y de información geográfica y proporciona una capa de citaciones lista para análisis; los indicadores completos de control de calidad se reportan en las Tablas 2-3.

Recibido: 29-11-2025. **Aceptado:** 09-02-2026. **Publicado:** 15-02-2026.

Cómo citar: Britto, A., Robledo, S., & Zuluaga, M. (2026). BibFusion: A Python package to integrate, deduplicate, and harmonize exported bibliographic records from Scopus and Web of Science for bibliometric analysis. *Iberoamerican Journal of Science Measurement and Communication; 6*(1), 1-22. DOI: 10.47909/ijsmc.342

Conclusiones/Valor. BibFusion ofrece un flujo de trabajo de integración reutilizable y auditable que reduce la inflación por duplicados y la fragmentación de metadatos, al tiempo que hace explícitas las decisiones de fusión y la incertidumbre residual para habilitar análisis posteriores transparentes.

PALABRAS CLAVE: bibliometría; cienciometría; integración entre bases de datos; Scopus; Web of Science; preprocesamiento de metadatos; desambiguación de autores; redes de citas; investigación reproducible.

ABSTRACT

Objective. The study presented BibFusion, a Python software package that harmonizes bibliographic exports from Scopus and Web of Science into a single, traceable, analysis-ready corpus for bibliometric and scientometric research.

Design/Methodology/Approach. BibFusion was capable of ingesting Scopus CSV and WoS TXT files, applying systematic normalization (e.g., ASCII/uppercase standardization of titles and SR keys, affiliation parsing with country extraction), and optionally enriching records via DOI-based resolution against OpenAlex to recover persistent identifiers (e.g., OpenAlex work IDs, ORCID when available, and OpenAlex author IDs). Cross-database integration employed a DOI-first deduplication cascade with a conservative fallback (title-year-first author) in the event that a DOI is absent. The authors were disambiguated through a canonical PersonID hierarchy (ORCID → OpenAlexAuthorID → normalized name). Citation strings were cleaned and remapped to ensure the preservation of consistent citation links, and journal/Scimago information was consolidated using ISSN/EISSN rules.

Results. In a demonstration on an entrepreneurial marketing query, BibFusion consolidated 436 source records into 253 unique main works and materialized a unified corpus of 8,569 articles. The resulting dataset demonstrated high levels of identifier and geographic completeness, and it provided an analysis-ready citation layer.

Conclusions/Value. BibFusion offers a reusable, auditable integration workflow that has been demonstrated to reduce duplicate inflation and metadata fragmentation. This workflow facilitates the explicit determination of merge decisions and residual uncertainty, thereby ensuring transparency in downstream analyses.

KEYWORDS: bibliometrics; scientometrics; cross-database integration; Scopus; Web of Science; metadata preprocessing; author disambiguation; citation networks; reproducible research.

1. INTRODUCCIÓN

Los análisis bibliométricos se utilizan ampliamente para rastrear la evolución de los dominios de investigación, pero la validez de estos hallazgos depende de la consistencia de los metadatos bibliográficos subyacentes (Zhang *et al.,* 2024). En la práctica, integrar registros provenientes de fuentes principales como Scopus y Web of Science sigue siendo un reto, ya que ambas plataformas difieren en su cobertura, en la estructura de exportación y en la completitud de los campos (Kumpulainen & Seppänen, 2022). Estas diferencias suelen generar ítems duplicados entre fuentes, variaciones en la escritura de los nombres de autores, afiliaciones ambiguas o incompletas y referencias no estandarizadas que fragmentan los enlaces de citación. Como resultado, los indicadores posteriores, como mapas por país e institución, tendencias temporales de producción y redes de coautoría o de citación, pueden sesgarse o volverse inestables si el preprocesamiento se realiza de forma *ad hoc*. Por ello, es esencial incorporar un paso de armonización reproducible antes de cualquier análisis específico del dominio (por ejemplo, búsquedas relacionadas con el marketing emprendedor), a fin de construir un corpus limpio e integrado que permita una cienciometría confiable (Nowakowska, 2025).

Integrar Scopus y Web of Science en un único corpus resulta difícil porque ambas bases difieren en cobertura, estructura de exportación y completitud de metadatos (Delgado-Quirós & Ortega, 2024). Los DOI pueden faltar o estar mal formados; los campos textuales (títulos

y referencias citadas) varían en mayúsculas, puntuación y normalización de caracteres; y los datos de autores y afiliaciones suelen ser heterogéneos o incompletos, lo que complica la atribución a nivel de personas y países (Visser *et al.,* 2021). En conjunto, estas discrepancias inflan duplicados, introducen ambigüedad en identidades de autor y fragmentan enlaces de citación, sesgando indicadores de productividad, colaboración y geografía científica (Nowakowska, 2025).

Para abordar este problema, presentamos BibFusion, un software reproducible y auditable que armoniza y fusiona exportaciones de Scopus y Web of Science en un conjunto de datos único, trazable y listo para el análisis. BibFusion estandariza metadatos, analiza afiliaciones (incluida la extracción de país) y, opcionalmente, enriquece registros mediante resolución por DOI en OpenAlex para recuperar identificadores persistentes de obras y autores cuando están disponibles (Nowakowska, 2025; Priem *et al.,* 2022). La integración aplica una desduplicación conservadora (DOI primero, con reglas de respaldo cuando falta) y consolida la autoría para soportar análisis consistentes (Visser *et al.,* 2021). El resultado es un corpus relacional en tablas vinculadas, acompañado de artefactos de auditoría que documentan procedencia y decisiones de fusión para facilitar revisión transparente y reproducibilidad (Maisano *et al.,* 2025).

El alcance del estudio queda definido por la consulta aplicada en Scopus y Web of Science, con énfasis en artículos de revista en inglés dentro de la ventana de publicación especificada para el corpus principal. Dado que la integración entre fuentes es más confiable cuando existen identificadores persistentes, BibFusion prioriza la coincidencia basada en DOI cuando está presente y recurre a reglas conservadoras de concordancia de metadatos cuando los DOI faltan o están mal formados. De manera similar, el enriquecimiento y la consolidación de autores se benefician de identificadores externos (p. ej., ORCID u OpenAlexAuthorID), que pueden ser incompletos entre fuentes; por ello, los casos no resueltos se conservan sin imponer decisiones forzadas y se visibilizan mediante salidas de auditoría para su revisión transparente. Es importante señalar que, aunque el corpus principal está acotado por la consulta, las referencias citadas pueden extenderse más allá de la ventana temporal y de las restricciones de tipo documental de la consulta, lo cual es esperable al construir redes de citación.

Este artículo aporta una canalización de preprocesamiento reproducible y lista para auditoría que armoniza exportaciones de Scopus y Web of Science en un corpus relacional unificado. El flujo de trabajo integra un enriquecimiento centrado en identificadores (mediante resolución de DOI en OpenAlex) con una desambiguación sistemática de autores a través de un PersonID canónico, y materializa el corpus fusionado como entidades vinculadas con procedencia explícita. Además de salidas listas para analizar productividad, colaboración y geografía científica, BibFusion genera artefactos de auditoría (p. ej., registros de alias, conflictos y fusiones) que hacen transparentes y reproducibles las decisiones de integración.

El resto del artículo se estructura de la siguiente manera. En la sección 1.1 se revisan trabajos relacionados sobre integración bibliográfica entre bases y desambiguación. En la sección 2 se describe el paquete, sus entradas/salidas y el modelo de datos unificado. En la sección 3 se detalla la metodología de preprocesamiento, incluyendo normalización, enriquecimiento basado en OpenAlex, reglas de desduplicación y fusión, desambiguación de autores, limpieza de citaciones y consolidación de información de revistas/SCImago. En la sección 4 se reporta el corpus resultante y los principales indicadores de control de calidad. En la sección 5 se discute implicaciones, limitaciones y orientaciones prácticas para su reutilización. Por último, en la sección 6 se concluye y plantea mejoras futuras. Los detalles complementarios del flujo de trabajo y la documentación se incluyen en los apéndices.

1.1 Revisión de la literatura

Los ecosistemas bibliométricos existentes ofrecen un soporte robusto para importar, mapear y visualizar exportaciones de Scopus y Web of Science, incluyendo paquetes e interfaces ampliamente utilizados como bibliometrix/Biblioshiny (Aria & Cuccurullo, 2017; Maisano *et al.,* 2025), VOSviewer (van Eck & Waltman, 2010) y CiteSpace (Chen, 2006), así como herramientas orientadas a Python para acceso

y análisis de tendencias como pybliometrics (Rose & Kitchin, 2019) y ScientoPy (Ruiz-Rosero *et al.,* 2019). Aunque estas herramientas son altamente efectivas una vez que los datos han sido ingeridos, la integración entre bases suele requerir procesamiento adicional para reconciliar esquemas de exportación heterogéneos e inconsistencias de metadatos (p. ej., DOI ausentes o registrados de forma inconsistente, variantes en las cadenas de autores y afiliaciones incompletas o no estandarizadas), especialmente cuando el objetivo es preservar y armonizar las referencias citadas entre fuentes. En respuesta, han surgido flujos de trabajo centrados en integración, desde *tosr,* que operacionalizó la integración Scopus-Web of Science (WoS) incorporando explícitamente información de referencias más allá de las fusiones estándar (Robledo *et al.,* 2024), hasta canalizaciones más recientes en Python como BibexPy, que enfatiza la fusión automatizada y el enriquecimiento de metadatos mediante servicios externos (Kara *et al.,* 2025), y kits configurables de emparejamiento como TeslaSCItoolkit, que formalizan el *record linkage* mediante métricas de similitud y clasificación de coincidencias (Nikolić *et al.,* 2024). En conjunto, la literatura previa refleja un desplazamiento hacia una integración centrada en identificadores y orientada a la automatización, pero persisten vacíos en (i) producir una representación relacional explícita, del tipo entidad-relación, que separe las entidades núcleo de las tablas de enlace, y (ii) proporcionar artefactos auditables y versionados que documenten las decisiones de fusión y desambiguación, en particular a nivel de referencias.

Los retos estructurales están ampliamente documentados en la integración bibliométrica y continúan dificultando la consolidación confiable de los registros de Scopus y Web of Science (McKay, 2026). En primer lugar, la desduplicación suele verse limitada por DOI ausentes, mal formados o registrados de manera inconsistente, lo que obliga a recurrir a reglas secundarias de coincidencia (p. ej., título-año-autor) e incrementa tanto los falsos positivos como los falsos negativos (Culbert *et al.,* 2025). En segundo lugar, la desambiguación de autores sigue siendo imperfecta debido a la variación ortográfica, el uso inconsistente de iniciales, las diferencias de transliteración (formas acentuadas vs. ASCII) y la homonimia, lo que puede fragmentar o fusionar indebidamente identidades y distorsionar los indicadores de productividad y colaboración (Kim & Owen-Smith, 2021). En tercer lugar, la información de referencias citadas es altamente heterogénea: las cadenas de referencia presentan con frecuencia formatos mixtos, puntuación irregular, años aislados o entradas mal formadas, lo cual rompe los enlaces de citación y debilita las redes de citación y cocitación, especialmente cuando se requiere integración a nivel de referencias entre fuentes (Cioffi *et al.,* 2022). Por último, los metadatos de revistas y afiliaciones suelen requerir normalización adicional, como, por ejemplo, la armonización de los nombres y abreviaturas de revistas, la corrección de afiliaciones incompletas y la incorporación de información de país ausente o implícita, lo que puede sesgar las medidas de colaboración y de geografía científica (Purnell, 2022). Estos desafíos motivan canalizaciones de preprocesamiento reproducibles que estandaricen los metadatos, fortalezcan la desduplicación, mejoren la trazabilidad y depuren los enlaces de citación antes del análisis bibliométrico y de redes.

A pesar del progreso sostenido en paquetes de software bibliométricos e importadores específicos por base de datos, persiste un vacío importante en la integración entre bases de datos: muchos flujos de trabajo priorizan los indicadores y la visualización posteriores, mientras que la capa de integración (incluida la armonización de referencias citadas) sigue siendo difícil de reproducir y auditar (Ng *et al.,* 2025). En particular, relativamente pocas aproximaciones formalizan la salida fusionada como un modelo entidad-relación explícita que separe entidades núcleo (p. ej., Articles, Authors, Journals, Affiliations) de tablas de enlace (p. ej., ArticleAuthor, Citation) y mantenga identificadores estables entre fuentes (Massari *et al.,* 2024). En consecuencia, pasos clave, como las decisiones de desduplicación, la resolución de identidad de autores y la limpieza de cadenas de citación, suelen implementarse de forma difícil de inspeccionar, validar o reejecutar de manera consistente cuando cambian los insumos. Además, los artefactos de auditoría (p. ej., listas de alias, banderas de conflicto y bitácoras de fusión) no se generan de forma sistemática, lo que obliga a los investigadores a depender de verificaciones manuales puntuales y de documentación

limitada (Elstad *et al.,* 2023). Esto motiva una canalización reproducible que produzca un corpus relacional listo para análisis, junto con archivos auditables que hagan transparentes las decisiones de integración y soporten la verificación sistemática y el refinamiento iterativo.

En este contexto, BibFusion se posiciona como una capa de integración reproducible y lista para auditoría entre Scopus y Web of Science que complementa los ecosistemas existentes de análisis bibliométrico. Aborda la brecha de integración materializando la salida fusionada como una estructura entidad-relación explícita y preservando los vínculos a nivel de referencias mediante tablas dedicadas de citación y de enlace (p. ej., claves SR/SR_ref estandarizadas y una lista de aristas de citación depurada) (Delgado-Quirós & Ortega, 2025). Metodológicamente, BibFusion aplica una estrategia de emparejamiento con prioridad al DOI y respaldos conservadores y, cuando los DOI pueden resolverse, enriquece los registros mediante OpenAlex para recuperar identificadores persistentes de obras y autores (Chavarro *et al.,* 2025). Asimismo, asigna un PersonID canónico (ORCID → OpenAlexAuthorIDx → nombre normalizado) para fortalecer la continuidad a nivel de autor cuando la cobertura de identificadores permite un enlace confiable (Rehs, 2021). De forma importante, el flujo de trabajo externaliza la incertidumbre mediante salidas de auditoría (alias, conflictos potenciales y bitácoras de fusión), lo que habilita la inspección transparente y la corrección iterativa (Ornstein, 2025). El resultado es un corpus relacional unificado y listo para análisis, con procedencia explícita, que puede consumirse directamente en análisis bibliométricos, de colaboración, geográficos y de redes de citación.

2. VISIÓN GENERAL DEL PAQUETE Y MODELO DE DATOS

2.1. Propósito del paquete, entradas y salidas

BibFusion es un paquete modular de preprocesamiento que armoniza las exportaciones bibliográficas de Scopus y Web of Science en un único corpus trazable y listo para el análisis. Ingiere archivos crudos CSV de Scopus y TXT de WoS, aplica una normalización sistemática a campos clave de metadatos (p. ej., títulos, claves de referencia, afiliaciones e información de país) y enriquece los registros mediante OpenAlex para recuperar identificadores persistentes de obras y autores cuando están disponibles (Velez-Estevez *et al.,* 2023). A continuación, la canalización ejecuta una desduplicación con prioridad al DOI y respaldos conservadores (Kara *et al.,* 2025), asigna PersonID canónicos para la desambiguación de autores, limpia y remapea las cadenas de citación para preservar enlaces de citación consistentes y consolida la información de revistas y la relacionada con SCImago usando reglas basadas en ISSN/EISSN (Mischo *et al.,* 2024). El flujo de trabajo se implementa como módulos reutilizables —ingestión, normalización, enriquecimiento, emparejamiento/fusión, desambiguación de autores y procesamiento de citaciones/revistas—, de modo que los usuarios pueden ejecutarlo de extremo a extremo para generar el conjunto de datos completo o correr etapas individuales para soportar actualizaciones incrementales y un control de calidad focalizado.

2.2. Síntesis del flujo de trabajo

La Figura 1 resume el flujo de trabajo de extremo a extremo implementado por BibFusion. Las exportaciones CSV de Scopus y TXT de Web of Science se procesan en paralelo mediante etapas modulares: ingestión y canonicalización, normalización de metadatos (incluyendo el análisis de afiliaciones y la extracción de país) y, de forma opcional, enriquecimiento basado en DOI a través de OpenAlex (Culbert *et al.,* 2025). Posteriormente, los registros estandarizados se desduplican y se fusionan utilizando una cascada de emparejamiento con prioridad al DOI y respaldos conservadores. El corpus integrado se estructura después mediante desambiguación de autores basada en PersonID, limpieza y remapeo de cadenas de citación (para poblar la tabla Citation), y consolidación de información de revistas/SCImago usando reglas basadas en ISSN/EISSN. Las salidas se materializan como siete tablas relacionales de entidades y se acompañan de artefactos de auditoría que documentan alias, conflictos potenciales y decisiones de fusión. Los diagramas de proceso a nivel de función (incluyendo dataframes intermedios y funciones de transformación) se presentan en el Apéndice A (Figuras A1-A2).

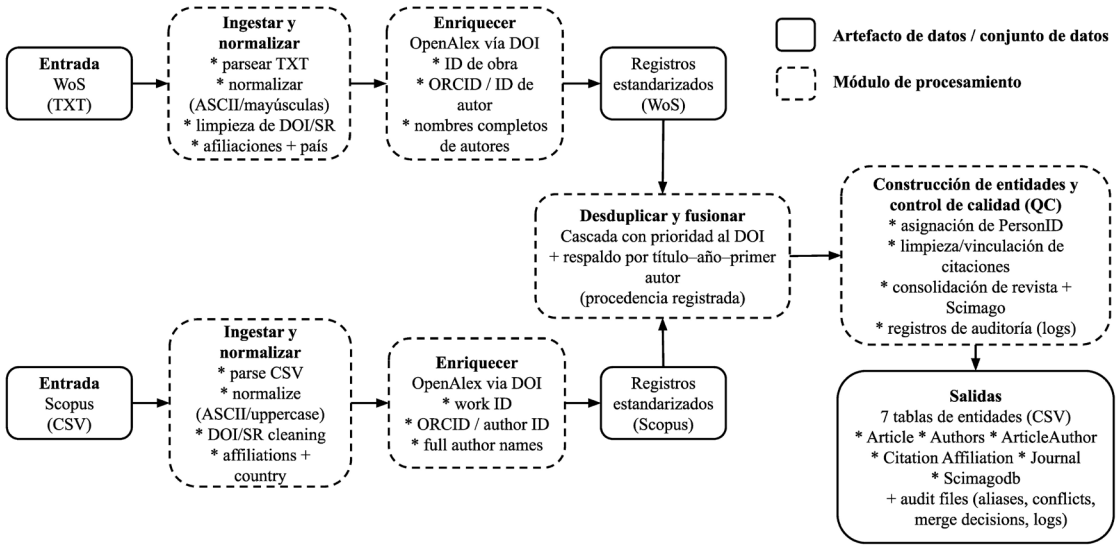

Figura 1. Visión general de la canalización de preprocesamiento. **Nota:** Elaboración propia.

2.3. Modelo de datos

La Figura 2 presenta el modelo entidad-relación (ER) unificado del corpus integrado producido por BibFusion. La canalización primero mapea las exportaciones de Web of Science y Scopus en tablas de *staging* específicas por fuente que comparten la misma estructura de entidades y las mismas convenciones de claves (un esquema de *staging* compartido e isomórfico), tras el proceso de parseo y normalización. Estandarizar la estructura en esta etapa —en lugar de fusionar directamente exportaciones heterogéneas— reduce las diferencias en nombres de

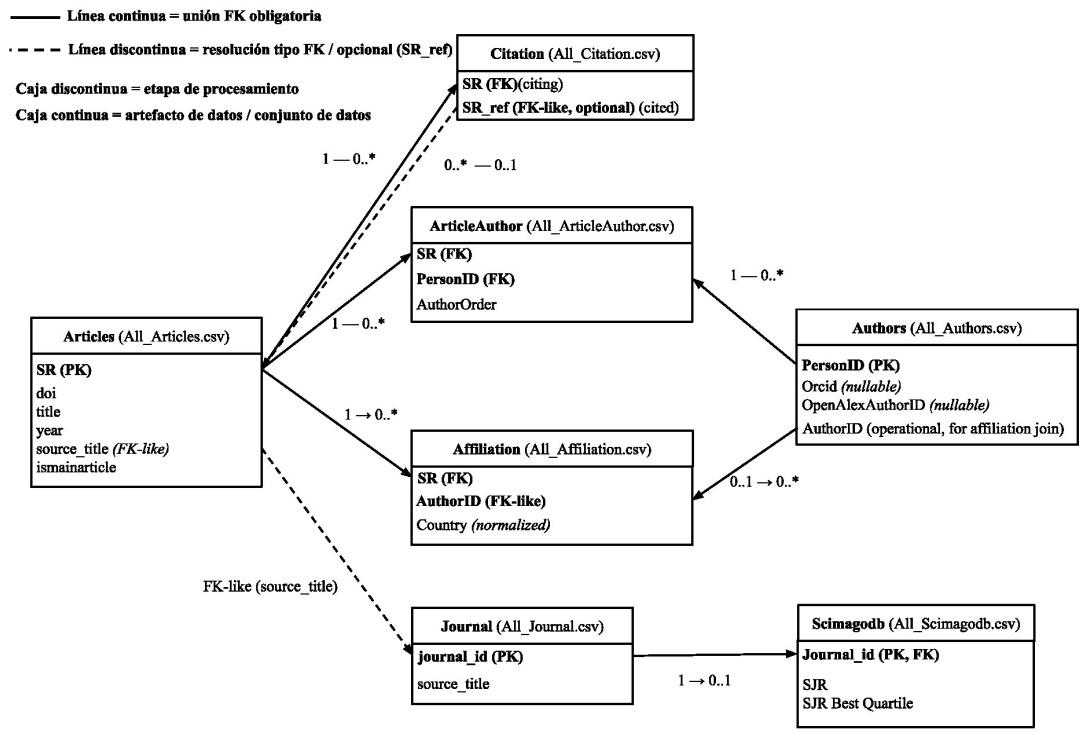

Figura 2. Modelo entidad-relación (ER) unificado del corpus integrado de BibFusion. **Nota:** Elaboración propia.

campos y formatos, y permite una desduplicación e integración consistentes basadas en reglas (Rehs, 2021). El corpus final se almacena como siete entidades relacionales vinculadas — Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal y Scimagodb. En la Figura 2, los conectores sólidos representan uniones (joins) obligatorias, mientras que los conectores punteados indican resoluciones de tipo FK u opcionales cuando la cobertura es incompleta (por ejemplo, SR_ref en Citation no siempre puede resolverse hacia un registro completo en Articles). La procedencia se preserva mediante indicadores de fuente (p. ej., sources_merged) y banderas de artículo principal vs. referencia (p. ej., ismainarticle), lo que habilita análisis bibliométricos, de colaboración, geográficos y de redes de citación reproducibles, con un filtrado transparente por tipo de registro (Lastilla *et al.,* 2022).

2.4. Inventario de salidas

La Tabla 1 enumera las siete salidas en formato CSV producidas por BibFusion y sus funciones dentro del esquema relacional unificado: Articles (registros a nivel de obra), Authors (identidades canónicas indexadas por PersonID), ArticleAuthor (vínculos obra-autor), Citation (aristas dirigidas de citación), Affiliation (instancias obra-autor-afiliación con país), Journal (revistas normalizadas indexadas por journal_id) y Scimagodb (métricas a nivel de revista, vinculables mediante journal_id). Para cada archivo, la Tabla 1 reporta la granularidad de las filas, la(s) clave(s) primaria(s) y los campos de unión que preservan la integridad relacional (p. ej., SR, PersonID, journal_id y los enlaces SR→SR_ref, señalando que SR_ref no siempre se resuelve hacia un registro completo en Articles). En conjunto, la tabla ofrece una guía de referencia concisa para ensamblar análisis reproducibles de tendencias, colaboración, geografía científica y redes de citación a partir de las entidades estandarizadas.

Además de las tablas de entidades resumidas en la Tabla 1, BibFusion organiza las salidas en tres directorios que separan los artefactos intermedios a nivel de fuente del corpus integrado final. WoS_results/ y Scopus_results/ almacenan salidas de *staging* estandarizadas y bitácoras específicas por fuente, generadas durante la ingestión, la normalización y (cuando se habilita) el enriquecimiento con OpenAlex, lo que facilita reejecuciones parciales y la depuración focalizada a nivel de cada fuente. Los entregables finales, listos para análisis, se escriben en all_data_wos_scopus/, que contiene las siete entidades relacionales en CSV junto con archivos de auditoría que documentan alias, conflictos potenciales y decisiones de fusión. Esta separación favorece una ejecución reproducible, manteniendo los productos intermedios claramente diferenciados del conjunto consolidado utilizado en los análisis posteriores. Para una reutilización transparente, se distribuye en el repositorio de BibFusion un diccionario de datos completo y versionado (definiciones de variables, procedencia y reglas de transformación): https://pypi.org/project/bibfusion.

3. METODOLOGÍA

3.1. Recolección de datos, exportaciones y estrategia de búsqueda

BibFusion ingiere dos fuentes bibliográficas primarias: exportaciones de Scopus en formato CSV y de Web of Science en formato texto plano (TXT). Los registros se obtienen mediante la interfaz estándar de exportación de cada plataforma e incluyen metadatos centrales (p. ej., título, autores, afiliaciones, año de publicación, información de la fuente/revista, DOI cuando está disponible, referencias citadas y conteos de citación). El uso de ambas fuentes amplía la cobertura, al tiempo que preserva la procedencia por fuente para asegurar la trazabilidad, permitiendo que la canalización armonice registros que se refieren a una misma publicación a través de bases de datos. Operativamente, BibFusion aplica parsers específicos por fuente y mapea ambas exportaciones a un esquema de *staging* idéntico (tablas alineadas por fuente), de modo que los pasos posteriores de normalización, enriquecimiento y desduplicación/fusión operen sobre representaciones armonizadas, en lugar de formatos crudos heterogéneos. Todas las exportaciones se generaron con la configuración de “registro completo” (full record), incluyendo las referencias citadas (véase Apéndice B).

CSV de salida	Granularidad de filas	Clave primaria (PK)	Claves foráneas / vínculos (cómo unir)	Columnas clave (ejemplos)	Propósito
All_Articles.csv	1 fila por registro de obra (registros principales + registros tipo referencia) después de la armonización/ fusión.	SR (canonical work/ reference key); doi as persistent identifier when present (recommended guard: SR + doi for rare SR collisions among references)	SR is referenced by All_ArticleAuthor and All_Citation; journal linkage via source_title / journal → All_Journal	SR, title, year, doi, ismainarticle, sources_merged, country, source_title, journal, cited_by, cited_reference_count, link	Tabla canónica de "obras" para análisis de tendencias/productividad y como lista de nodos para redes de colaboración y citación.
All_Authors.csv	1 fila por identidad única de persona después de la desambigüación.	PersonID	PersonID is referenced by All_ArticleAuthor	PersonID, AuthorFullName, AuthorName, Orcid, OpenAlexAuthorID, ResearcherID, Email, AuthorID	Tabla canónica de identidades de autor que permite análisis a nivel de autor y uniones consistentes entre salidas.
All_ArticleAuthor.csv	1 fila por vínculo de autoría (asociación persona-obra).	(SR, PersonID, AuthorOrder)	SR → All_Articles.SR; PersonID → All_Authors.PersonID	SR, PersonID, AuthorOrder, CorrespondingAuthor, OpenAlexAuthorID, AuthorID, openalex_work_id	Implementa la relación de autoría muchos-a-muchos (redes de coautoría, análisis de orden de autor, productividad por autor).
All_Citation.csv	fila por arista de citación dirigida (citante → citado).	(SR, SR_ref)	SR → All_Articles.SR (citing work); SR_ref → All_Articles.SR when available (optional mapping; some cited items remain external)	SR, SR_ref	Lista de aristas para redes de citación/ cocitación y enlace de referencias depurado entre fuentes.
All_Affiliation.csv	1 fila por instancia (obra-autor-afiliación).	(SR, AuthorID, Affiliation) (Country is derived and can be retained as an attribute)	SR → All_Articles.SR; AuthorID can link to All_Authors.AuthorID and/or to All_ArticleAuthor (then to PersonID)	SR, AuthorID, Affiliation, Country	Mapeo institucional y geográfico (afiliaciones/países), que soporta colaboración por país y análisis basados en afiliación.
All_Journal.csv	1 fila por revista/ fuente normalizada.	journal_id	Join from All_Articles using source_title (preferred) and/or journal → All_Journal.source_title/journal; journal_id → All_Scimagodb.journal_id	journal_id, source_title, journal	Metadatos de revistas normalizados que habilitan análisis por outlet y sirven de puente hacia métricas tipo Scimago.
All_Scimagodb.csv	1 fila por revista con métricas disponibles.	journal_id	journal_id → All_Journal.journal_id	journal_id, Title, Issn, eIssn, SCImago Journal Rank (SJR), SJR Best Quartile, H index, Publisher, Country, Categories, Areas, Coverage	Métricas contextuales a nivel de revista (cuartiles/SJR/índice h, categorías) para benchmarking de outlets y análisis estratificados.

Tabla 1. Inventario de salidas de BibFusion y claves de unión relacional. Nota: SR_ref es una clave normalizada de referencia citada y no siempre se resuelve hacia un registro completo en All_Articles (p. ej., libros, informes u obras fuera de cobertura). En esos casos, la arista de citación se conserva para análisis de redes de citación utilizando la estructura SR → SR_ref. PersonID es el identificador canónico de autor utilizado en todas las salidas, asignado mediante una regla de prioridad (ORCID → OpenAlexAuthorID → nombre normalizado) y puede completarse mediante enriquecimiento aun cuando esté ausente en las exportaciones crudas. Datos generados por BibFusion.

Los registros se recuperaron usando criterios de selección alineados con Scopus y Web of Science para maximizar la comparabilidad entre fuentes. En ambas bases, se buscó la frase exacta "entrepreneurial marketing" en el campo de Título (Title) para priorizar la precisión y reducir el ruido temático frente a búsquedas basadas en resúmenes o en palabras clave. Los resultados se restringieron a artículos de revista en inglés (tipo de documento: Article) publicados entre 2005 y 2025 (inclusive). La consulta en Scopus se implementó mediante la sintaxis de búsqueda avanzada (operacionalizada como PUBYEAR > 2004 AND PUBYEAR < 2026, con DOCTYPE "ar" correspondiente a Article), mientras que la consulta en Web of Science se implementó usando etiquetas de campo de la Core Collection (p. ej., TI para Título y PY para año de publicación). En Web of Science, PY puede reflejar el año de publicación final o, cuando aplica, el de early access; por ello, BibFusion preserva los campos de año específicos de la fuente, a la vez que estandariza las salidas integradas para soportar reportes posteriores consistentes. Las búsquedas y las exportaciones se ejecutaron el 16 de enero de 2026.

Scopus (búsqueda avanzada):

```
TITLE("entrepreneurial marketing") AND PUBYEAR > 2004 AND PUBYEAR < 2026 AND (LIMIT-TO(DOCTYPE,"ar")) AND (LIMIT-TO(LANGUAGE,"English"))
```

Web of Science Core Collection (búsqueda avanzada):

```
TI=("entrepreneurial marketing") AND PY=(2005-2025) AND LA=(English) AND DT=(Article)
```

BibFusion está diseñado para ejecutarse sobre exportaciones de registros completos tanto de Scopus como de Web of Science. En consecuencia, el flujo de trabajo asume que (i) los registros de Scopus se exportan en formato CSV con la opción "Select all information" habilitada y (ii) los registros de Web of Science se exportan en formato texto plano (TXT) usando "Full Record and Cited References". Estas configuraciones garantizan la disponibilidad de la información mínima requerida para la armonización e integración: descriptores bibliográficos (título, año, tipo de documento, identificadores de la fuente/revista como ISSN/EISSN y, cuando esté disponible, la paginación), identificadores persistentes (en especial el DOI), campos de autoría y afiliación/dirección (para el enlace de autoría y la extracción del país) y cadenas completas de referencias citadas (para construir y normalizar claves SR/SR_ref y poblar la entidad Citation). Para evitar la pérdida de información, debe deshabilitarse cualquier opción de truncamiento destinada a la compatibilidad con hojas de cálculo (p. ej., "optimize for Excel"), ya que puede acortar campos largos como afiliaciones, resúmenes y referencias citadas y comprometer el emparejamiento y la construcción de enlaces de citación. Las configuraciones de exportación utilizadas en este estudio se documentan en el Apéndice B (Figuras B1-B2).

BibFusion se apoya en exportaciones que, en conjunto, cubren cinco componentes de metadatos. (i) Metadatos bibliográficos núcleo — título, año de publicación e información de la fuente/revista (incluyendo ISSN/EISSN cuando esté disponible), junto con volumen/número/páginas— permiten la normalización y la consolidación de los outlets. (ii) Identificadores persistentes y específicos de base de datos, especialmente el DOI, constituyen el ancla principal para el emparejamiento y la desduplicación entre fuentes. (iii) Campos de autoría y de afiliación/dirección habilitan el enlace de autoría, el parseo institucional y la extracción del país para análisis geográficos. (iv) Cadenas completas de referencias citadas alimentan la entidad Citation y respaldan la normalización SR/SR_ref para preservar enlaces de citación consistentes. (v) Descriptores adicionales (p. ej., conteos de citación, indicadores de acceso abierto y palabras clave/resúmenes) se conservan para mantener la trazabilidad y para soportar análisis descriptivos y temáticos posteriores una vez generado el corpus unificado.

3.2. Consideraciones de calidad de datos y limitaciones

Dado que BibFusion requiere exportaciones completas de Scopus y Web of Science, la canalización está diseñada para manejar faltantes e inconsistencias del mundo real que persisten incluso en configuraciones de "registro completo".

Entre los problemas más comunes se incluyen: (i) campos de metadatos vacíos o parcialmente diligenciados (especialmente afiliaciones, direcciones e identificadores); (ii) DOI ausentes, mal formados o con formatos inconsistentes, lo que puede impedir coincidencias exactas entre fuentes; y (iii) heterogeneidad en las cadenas de autores y de fuentes, incluyendo variaciones en puntuación, orden, iniciales y diacríticos (formas acentuadas vs. ASCII). BibFusion aborda estos retos mediante normalización sistemática (estandarización en ASCII y en mayúsculas para campos textuales clave, como títulos y claves SR), limpieza y normalización de DOI cuando están presentes, y una cascada de emparejamiento conservadora que prioriza la alineación exacta por DOI y aplica un respaldo ligero basado en título-año-primer autor solo cuando el DOI no está disponible. En lugar de descartar registros con campos incompletos, la canalización los conserva siempre que sea posible construir una clave canónica de obra/referencia, preservando la procedencia y reduciendo el riesgo de fusiones erróneas. Las identidades de autor se consolidan mediante un *PersonID* canónico (ORCID → OpenAlexAuthorID → nombre normalizado), y las cadenas de referencias citadas se depuran para eliminar entradas mal formadas o de baja información (p. ej., años aislados) que, de otro modo, romperían el enlace de citaciones. Las ambigüedades residuales no se resuelven de forma silenciosa: BibFusion genera artefactos de auditoría (alias, conflictos potenciales y bitácoras de fusión) para que los casos inciertos puedan revisarse y corregirse sin comprometer la reproducibilidad.

3.3. Definición del corpus

La unidad de análisis en BibFusion es el registro bibliográfico que figura en el corpus integrado. Desde el inicio, distinguimos dos tipos de registro: registros principales, que corresponden a publicaciones recuperadas directamente de Scopus y Web of Science mediante la consulta y los filtros especificados, y registros de referencia, que corresponden a ítems citados extraídos (parsed) de las listas de referencias de los registros principales. Ambos se almacenan en la entidad Articles para mantener un único espacio de nodos para la construcción de redes, pero difieren sustancialmente en la completitud: los registros principales suelen contener metadatos ricos e identificadores persistentes (p. ej., DOI y campos de fuente), mientras que los registros de referencia pueden contener únicamente una clave de referencia depurada/normalizada (SR/SR_ref) y un conjunto limitado de atributos. Las relaciones de citación dirigidas se representan en la entidad *Citation* como aristas desde un registro principal citante (SR) hacia una clave de referencia citada (SR_ref), y no todos los ítems citados necesariamente se mapean a un registro de artículo completamente descrito dentro del conjunto de datos. Las relaciones de autoría se capturan en *ArticleAuthor* solo cuando hay suficiente metadato para establecer un enlace confiable a un *PersonID* canónico. Esta representación soporta análisis centrados en el corpus recuperado (p. ej., indicadores de productividad y geografía científica sobre registros principales) y, al mismo tiempo, habilita análisis de redes de citación que incorporan de forma transparente nodos "solo-referencia" sin sobrestimar su calidad como metadatos.

Usando la estrategia de búsqueda de la Sección 3.1, la tabla integrada *Articles* contiene 8.569 registros bibliográficos en total. De estos, 253 son registros principales recuperados directamente de Scopus y Web of Science (ismainarticle = True), mientras que 8.316 son registros de referencia extraídos de cadenas de referencias citadas (ismainarticle = False). La procedencia se conserva en *sources_merged:* 5.178 registros se observan en ambas fuentes (Scopus y Web of Science), 2.541 son exclusivos de Scopus y 850 son exclusivos de Web of Science. En el corpus principal, específicamente, 183 registros se emparejan entre ambas fuentes, 59 son exclusivos de Scopus y 11 exclusivos de Web of Science. La disponibilidad de DOI es alta en general (8.054/8.569 = 94,0%) y se mantiene sustancial en los registros principales (220/253 = 87,0%); en particular, la cobertura de DOI entre los registros principales emparejados en ambas fuentes alcanza 98,9%, lo que refleja la complementariedad de ambas bases para la recuperación de identificadores. El corpus principal abarca 2005-2025, tal como lo define la consulta, mientras que los registros de referencia se extienden más allá de esta ventana (p. ej., desde 1900 hasta 2026) porque heredan años de los ítems citados y pueden incluir, ocasionalmente, valores de año de *early access* o valores ruidosos.

3.4 Canalización BibFusion y reproducibilidad

Con base en el flujo de trabajo resumido en la Figura 1, BibFusion implementa una canalización modular de procesamiento de datos de extremo a extremo que transforma exportaciones crudas de Scopus (CSV) y Web of Science (TXT) en un corpus relacional armonizado, desduplicado y auditable, adecuado para análisis bibliométricos y de redes reproducibles.

BibFusion inicia con la ingestión de exportaciones CSV de Scopus y TXT de Web of Science mediante parsers específicos por fuente, que mapean campos crudos heterogéneos a un esquema común de staging, conservando al mismo tiempo la procedencia por fuente para asegurar la trazabilidad. Luego se aplica una capa uniforme de normalización para estabilizar el emparejamiento posterior: los campos textuales clave (p. ej., títulos y cadenas de autor) se convierten en una representación consistente en ASCII y en mayúsculas, y se recortan para eliminar espacios espurios y artefactos de formato. En paralelo, las claves de referencia se estandarizan mediante limpieza y normalización de las cadenas *SR/SR_ref* (p. ej., removiendo puntuación inconsistente y tokens de baja información) para soportar un enlace de citaciones y una desduplicación más robustos. Finalmente, los campos de afiliación/dirección se parsean para derivar un atributo de país normalizado; las menciones repetidas o inconsistentes de país dentro de un registro se consolidan de modo que los indicadores a nivel de país y de geografía científica permanezcan comparables entre Scopus y WoS tras la integración.

Después de la normalización, BibFusion puede enriquecer opcionalmente los registros mediante la resolución basada en DOI contra OpenAlex para recuperar identificadores persistentes y mejorar el enlace entre fuentes. Cuando existe un DOI válido, la canalización consulta OpenAlex para obtener un identificador estable de obra (p. ej., OpenAlex work ID) y para ampliar los metadatos de autor, incluyendo OpenAlexAuthorID y ORCID cuando esté disponible, lo que posteriormente fortalece la desambiguación de autores y las uniones entre entidades. Este paso de enriquecimiento es deliberadamente conservador y no bloqueante: los registros con DOI ausentes o inválidos, o cuyos DOI no pueden resolverse, se omiten y se registran en bitácoras, en lugar de producir fallos; así, los registros incompletos permanecen en el corpus integrado y los identificadores no resueltos y los resultados del enriquecimiento quedan documentados de forma transparente para revisión.

BibFusion integra registros de Scopus y Web of Science mediante una estrategia de desduplicación y fusión con prioridad al DOI, aplicando coincidencia exacta sobre DOI normalizados cuando están disponibles y usando un respaldo conservador basado en la concordancia título-año-primer autor cuando el DOI falta. Cuando dos registros se fusionan, la canalización sigue una política transparente de "mejor registro", reteniendo los valores más informativos en los campos superpuestos, mientras que preserva la procedencia mediante sources_merged y banderas relacionadas que indican el origen de cada fila integrada. Luego, el corpus unificado se materializa en las entidades relacionales definidas en la Sección 2: los autores se desambiguan usando un *PersonID* canónico (ORCID → OpenAlexAuthorID → nombre normalizado) y se propagan a Authors y ArticleAuthor; las cadenas de citación se limpian y se estandarizan como aristas dirigidas en *Citation;* y los metadatos de outlets se consolidan en *Journal* y se vinculan con *Scimagodb* mediante reglas de normalización basadas en ISSN/EISSN y en título.

BibFusion incorpora el control de calidad a lo largo del flujo para reducir el ruido sin ocultar la incertidumbre. Entradas de referencia mal formadas o de baja información (p. ej., claves SR/SR_ref vacías, años aislados o cadenas de nopuntuación de informativas) se filtran durante la limpieza de citaciones, mientras que situaciones potencialmente ambiguas —como coincidencias limítrofes en desduplicación, colisiones de nombres de autor o resolución incompleta de identificadores— se registran y se visibilizan, en lugar de forzar silenciosamente una única interpretación. Además de las siete tablas de entidades, la canalización genera artefactos de auditoría (p. ej., listas de alias, archivos de revisión de conflictos y bitácoras de fusión/enriquecimiento) que proveen un rastro explícito de decisiones clave y de casos que requieren inspección. BibFusion también soporta diagnósticos reproducibles resumiendo indicadores

centrales de control de calidad —como duplicados eliminados bajo la cascada de emparejamiento, cobertura de DOI (global y para registros principales), completitud de extracción de país para registros principales, y proporción de búsquedas de enriquecimiento no resueltas— para que los usuarios evalúen la calidad antes de ejecutar análisis bibliométricos, de colaboración, geográficos o de redes de citación.

BibFusion se distribuye como un paquete de Python versionado con dependencias explícitas especificadas en requirements.txt, lo que habilita la recreación del entorno y ejecución consistente entre sistemas. La canalización se ejecuta a través de un único punto de entrada (p. ej., python run_main.py) y sigue una convención fija de directorios de entrada/salida que separa los artefactos de staging por fuente del conjunto de datos final integrado (Scopus_results/, WoS_results/ y all_data_wos_scopus/). La reproducibilidad se refuerza mediante salidas deterministas (las siete tablas de entidades más los artefactos de auditoría) y archivos de log que registran las operaciones de enriquecimiento y fusión. En términos computacionales, el tiempo de ejecución escala principalmente con el número de registros procesados y con el enriquecimiento opcional vía OpenAlex; la resolución basada en DOI suele ser el principal cuello de botella y puede verse limitada por restricciones externas de tasa (rate limits) de la API, por lo que las ejecuciones a gran escala pueden requerir batching, caching o ejecución programada para completarse de manera confiable. La versión utilizada en este estudio es BibFusion v1.0.0, disponible en el repositorio del proyecto: https://github.com/ladmepaz/bibfusion. El repositorio incluye el código fuente, un ejemplo de estructura de carpetas y un diccionario de datos versionado (CSV/MD) que documenta todas las variables de salida, incluyendo definiciones, procedencia (Scopus/WoS/derivada) y reglas de normalización.

4. RESULTADOS

4.1. Corpus integrado producido por BibFusion

BibFusion produce un corpus relacional unificado organizado en siete entidades —Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal y Scimagodb— según lo define el modelo de datos en la Sección 2.3. La Tabla 2 presenta los tamaños finales de cada entidad para la ejecución descrita en la Sección 3.1, proporcionando un resumen compacto del conjunto de datos integrado generado por la canalización.

Entidad (CSV)	Filas	Lo que representa
Articles (All_Articles.csv)	8,569	Registros de obras (registros principales + registros de referencia)
Authors (All_Authors.csv)	17,219	Personas desambiguadas (PersonID único)
ArticleAuthor (All_ArticleAuthor.csv)	28,254	Vínculos de autoría (asociaciones obra-persona)
Citation (All_Citation.csv)	24,392	Aristas de citación dirigidas (SR → SR_ref)
Affiliation (All_Affiliation.csv)	34,488	Instancias obra-autor-afiliación (con país)
Journal (All_Journal.csv)	831	Revistas/fuentes normalizadas (journal_id)
Scimagodb (All_Scimagodb.csv)	740	Revistas con métricas (vinculables por journal_id)

Tabla 2. Tamaños de las entidades del corpus unificado generado por BibFusion.
Nota: La entidad Articles incluye 253 registros principales y 8.316 registros de referencia extraídos de referencias citadas. Datos generados por los autores.

Para la consulta de *entrepreneurial marketing,* la entidad *Articles* contiene 8.569 registros bibliográficos en total, compuestos por 253 registros principales recuperados directamente de Scopus/Web of Science y 8.316 registros de referencia extraídos de cadenas de referencias citadas. La capa de autoría está conformada por 17.219 personas desambigüadas en *Authors* y 28.254 vínculos de autoría en *ArticleAuthor,* mientras que la capa de citación contiene 24.392 aristas dirigidas de citación en *Citation*. Las entidades restantes (Affiliation, Journal, Scimagodb) capturan la información institucional y geográfica estandarizada, así como los metadatos y métricas de los *outlets* consolidados, que contextualizan el corpus unificado para los análisis posteriores.

4.2. Resultados de integración y preparación del conjunto de datos (evidencia de control de calidad)

La Tabla 3 presenta indicadores clave de integración y calidad que demuestran que el corpus queda listo para el análisis tras la armonización. A nivel de los registros principales (es decir, publicaciones recuperadas directamente por la consulta), BibFusion preserva la procedencia entre fuentes mediante *sources_merged,* mostrando que 183/253 (72,3%) de los registros principales están presentes tanto en Scopus como en Web of Science, mientras que 59/253 (23,3%) son exclusivos de Scopus y 11/253 (4,3%) son exclusivos de Web of Science. Este solapamiento implica que, sin integración, la recuperación cruda combinada ascendería a 436 registros principales específicos por base (Scopus: 242, Web of Science: 194), con 183 duplicados entre fuentes; BibFusion consolida estos en 253 obras principales únicas, evitando el doble conteo en indicadores posteriores de productividad, colaboración y geografía científica.

Indicador	Valor
Registros principales únicos (integrados)	253
Registros de referencia (extraídos de las citaciones)	8,316
Registros principales por procedencia (sources_merged)	183 ambos; 59 solo Scopus; 11 solo Web of Science
Recuperación cruda de registros principales antes de la fusión (Scopus + WoS)	242 + 194 = 436
Duplicados entre fuentes consolidados (registros principales)	183
Cobertura de DOI (todos los artículos en Articles)	8,054 / 8,569 = 94.0%
Cobertura de DOI (registros principales)	220 / 253 = 87.0%
Cobertura de DOI (registros principales emparejados en ambas fuentes)	181 / 183 = 98.9%
Completitud de país en Articles (registros principales))	252 / 253 = 99.6%
Filas de afiliación con país diligenciado	34,488 / 34,488 = 100%
Aristas de citación (únicas)	24,392 (0 duplicate edges)
Aristas de citación originadas en registros principales	23,927 / 24,392 = 98.1%
Extremo citante resoluble en Articles (SR)	24,392 / 24,392 = 100%
Extremo citado resoluble en Articles (SR_ref)	13,275 / 24,392 = 54.4%
Revistas vinculadas a métricas de Scimagodb (journal_id)	727 / 831 = 87.5%
Cobertura temporal	Principales: 2005-2025; Referencias: 1900-2026

Tabla 3. Resultados de integración e indicadores de calidad que respaldan la preparación del conjunto de datos para el análisis. **Nota:** Se espera que la tasa de resolubilidad del extremo citado sea <100% porque muchos ítems citados (p. ej., libros o fuentes fuera de cobertura) no aparecen como registros completamente descritos en *All_Articles;* las aristas se conservan utilizando la clave *SR_ref* normalizada. Datos generados por el software en Python.

La completitud de los identificadores y metadatos respalda adicionalmente la preparación del conjunto de datos. La cobertura de DOI es de 94,0% en general (8.054/8.569) y de 87,0% para los registros principales (220/253), con un fuerte efecto de complementariedad entre fuentes: la cobertura de DOI entre los registros principales emparejados alcanza 98,9% (181/183), mientras que es menor en los registros principales exclusivos por fuente (solo Scopus: 57,6%; solo Web of Science: 45,5%), lo que indica que la fusión mejora sustancialmente la disponibilidad de identificadores en el corpus unificado. Los metadatos geográficos son altamente completos para el corpus principal: el campo de país a nivel de artículo está poblado para 252/253 (99,6%) registros principales, y la entidad *Affiliation* provee un valor de país normalizado para el 100% de las filas de afiliación, con afiliaciones disponibles para 251/253 (99,2%) registros principales. La consolidación de outlets también es sólida: 87,5% de las revistas normalizadas (727/831) están vinculadas a métricas de *Scimagodb* mediante *journal_id,* habilitando análisis estratificados por indicadores de revista.

Por último, el enlace de citaciones exhibe una estructura confiable para la construcción de redes y mantiene la transparencia respecto a los límites de cobertura. La tabla *Citation* contiene

24.392 aristas dirigidas, sin aristas duplicadas, bajo la clave (SR, SR_ref); todas las claves citantes (SR) se resuelven a la entidad *Articles* (100%), y 98,1% de las aristas de citación se originan en registros principales. Para el extremo citado, 54,4% de los valores *SR_ref* se mapean a un registro en *Articles,* mientras que el resto corresponde a ítems fuera de cobertura (p. ej., libros o fuentes que no están representadas como registros completos). Esto produce una red de citación que es a la vez utilizable (aristas preservadas y estandarizadas) y explícita sobre qué nodos citados son claves "solo-referencia" en lugar de obras completamente descritas.

5. DISCUSIÓN

5.1. Implicaciones

La evidencia de integración y de control de calidad reportada en las Tablas 2-3 tiene implicaciones directas para la validez de los análisis bibliométricos posteriores. En primer lugar, consolidar duplicados entre fuentes y preservar la procedencia (sources_merged) evita el doble conteo y estabiliza los indicadores de productividad y de tendencias; sin desduplicación, una misma obra puede aparecer como dos registros distintos, inflando los conteos de publicaciones y distorsionando los patrones temporales y los rankings de los outlets. En segundo lugar, la mejora en la cobertura de identificadores —en particular, la disponibilidad casi completa de DOI entre los registros emparejados entre Scopus y Web of Science— fortalece el enlace entre registros y habilita uniones más confiables con recursos externos, reduciendo la fragmentación en los flujos de trabajo de citación y de enriquecimiento de metadatos. En tercer lugar, la alta completitud de la extracción de país para los registros principales (y la población total de país en la tabla Affiliation) mejora sustancialmente los análisis de geografía científica al reducir el ruido de "país desconocido" y permitir que los mapas de producción y colaboración a nivel de país reflejen señales institucionales, en lugar de artefactos por datos faltantes. Por último, el *PersonID* canónico y la tabla explícita de vínculos de autoría permiten construir redes de coautoría y colaboración más creíbles al reducir la fragmentación o la fusión indebidas de autores causadas por variantes ortográficas (p. ej., acentos, iniciales u ordenamiento). En conjunto, estas mejoras de control de calidad transforman el conjunto de datos de exportaciones heterogéneas en un corpus relacional trazable, donde los indicadores bibliométricos, las comparaciones por país y las medidas de redes pueden calcularse de forma reproducible, con menor riesgo de sesgos derivados de inconsistencias en los metadatos (Crystal-Ornelas *et al.,* 2022).

5.2. Posicionamiento y limitaciones

Conceptualmente, BibFusion se sitúa entre los *toolkits* bibliométricos de propósito general y los scripts *ad hoc,* específicos de proyecto, orientados a la integración. Muchos flujos de trabajo existentes pueden importar exportaciones de Scopus y Web of Science en una sola tabla plana para indicadores descriptivos, pero con frecuencia dejan las referencias citadas como cadenas específicas de cada base de datos y no mantienen un espacio compartido y normalizado de claves de referencia entre fuentes; en consecuencia, la construcción de redes de citación y la desduplicación a nivel de referencias pueden ser frágiles y difíciles de reproducir. BibFusion aborda esta brecha materializando el corpus integrado como un modelo entidad-relación explícito y generando tablas de enlace dedicadas (ArticleAuthor, Citation, Affiliation) ancladas a identificadores canónicos (SR y PersonID). La procedencia se preserva mediante sources_merged, y las decisiones de integración se externalizan a través de artefactos de auditoría para facilitar inspección y reejecuciones. En la práctica, este diseño reduce el doble conteo entre fuentes, estabiliza la atribución a nivel de autor y país para el corpus principal, y produce una lista de aristas de citación lista para análisis con límites claros y transparentes entre ítems citados resolubles y referencias fuera de cobertura.

Al mismo tiempo, BibFusion no elimina las limitaciones inherentes a los metadatos bibliográficos (Visser *et al.,* 2021). Dado que la integración entre fuentes se ancla en identificadores persistentes, una estrategia con prioridad al DOI sigue siendo vulnerable a DOI ausentes, mal formados o registrados de manera inconsistente; aunque BibFusion aplica reglas de respaldo conservadoras cuando el DOI no está

disponible, cualquier emparejamiento no basado en el DOI conserva un riesgo residual tanto de coincidencias perdidas como de fusiones falsas ocasionales. La desambiguación de autores se fortalece mediante identificadores ORCID y OpenAlex, combinados con la normalización de nombres, pero la homonimia y la cobertura incompleta de los identificadores pueden seguir produciendo casos ambiguos que ameritan revisión humana—de ahí la generación de archivos explícitos de conflictos y bitácoras de alias, en lugar de forzar decisiones determinísticas. De manera similar, el enlace de citaciones es necesariamente incompleto cuando los ítems citados quedan fuera de la cobertura de las bases exportadas (p. ej., libros, informes o fuentes no indexadas); BibFusion preserva estas relaciones como nodos identificados por clave de referencia, pero no puede garantizar metadatos completos para obras fuera de cobertura. En este sentido, la contribución de BibFusion no es "resolver" metadatos imperfectos, sino proveer un flujo de trabajo de integración reproducible y listo para auditoría, que reduce inconsistencias evitables y hace explícita la incertidumbre remanente para un análisis bibliométrico posterior riguroso.

5.3. Orientaciones prácticas

Las salidas relacionales de BibFusion están estructuradas para soportar análisis bibliométricos comunes de manera directa y reproducible. Para el análisis de tendencias y productividad, la tabla *Articles* funciona como la unidad primaria de análisis —por lo general, filtrada a *ismainarticle = True*—, de modo que las series temporales de publicaciones (por año) y los resúmenes a nivel de outlet (vía source_title/journal) pueden calcularse sobre el corpus desduplicado sin inflar los conteos; el campo *sources_merged,* además, permite realizar verificaciones de sensibilidad según la procedencia por fuente. Para geografía científica, los resultados por país pueden obtenerse en dos resoluciones complementarias: una vista ligera usando el campo de país a nivel de artículo en *Articles* para mapas de producción del corpus principal, y una vista más fina usando la tabla *Affiliation* para analizar vínculos autor-institución-país (p. ej., producción multinacional, patrones de colaboración institucional y enfoques de conteo fraccional) (Demaine, 2022; Hottenrott *et al.,* 2021; Matveeva *et al.,* 2022; Sivertsen *et al.,* 2025).

Para análisis de coautoría, *ArticleAuthor* codifica una estructura bipartita normalizada que vincula obras (SR) con autores desambiguados (PersonID). Esta representación permite derivar redes de coautoría a partir de relaciones autor-autor dentro de cada SR, con la opción de ponderar las aristas según el número de publicaciones compartidas. Para análisis basados en citaciones, *Citation* provee una lista de aristas dirigidas depurada (SR → SR_ref) que soporta grafos de citación, medidas básicas de impacto (grado de entrada/salida) y análisis de cocitación (Chen *et al.,* 2024; Wang *et al.,* 2023; Yang *et al.,* 2024). Cuando una clave citada (SR_ref) se resuelve a un registro en *Articles,* los nodos citados pueden enriquecerse con metadatos completos; cuando no se resuelve, la arista se conserva con la clave de referencia normalizada para que la estructura de la red permanezca intacta, dejando claramente marcados los ítems fuera de cobertura. Por último, Journal y Scimagodb contextualizan los resultados a nivel de outlet al vincular publicaciones con identidades de revista normalizadas y, cuando está disponible, con indicadores como cuartiles del SJR y categorías temáticas, habilitando reportes estratificados consistentes a través del corpus unificado (Lim *et al.,* 2024; Schmal, 2024; Vaccaro *et al.,* 2022).

Para obtener resultados consistentes, BibFusion debe ejecutarse sobre exportaciones completas de ambas bases usando configuraciones estandarizadas. En Scopus, exporte los registros como CSV con *"Select all information"* habilitado; en Web of Science, exporte como archivo de texto plano (TXT) con *"Full Record and Cited References"* seleccionado (Apéndice B). En ambos casos, debe evitarse cualquier opción de truncamiento orientada a la compatibilidad con hojas de cálculo (p. ej., "optimize for Excel"), ya que puede acortar campos largos, como afiliaciones, resúmenes y cadenas de referencias, lo que degrada el emparejamiento y el enlace de citaciones. También es recomendable mantener estable la estrategia de búsqueda entre fuentes (campos, años, idioma y tipo documental alineados) y archivar las consultas exactas y las fechas de exportación junto con los archivos crudos para soportar la reproducibilidad completa.

Durante la ejecución, una buena práctica es preservar la estructura de carpetas por defecto (Scopus_results/, WoS_results/, all_data_wos_scopus/) y conservar todas las bitácoras y los artefactos de auditoría generados. Se recomienda revisar las salidas de auditoría —en particular, archivos de alias/conflictos y bitácoras de fusión— antes de realizar análisis a nivel de autor o de institución, y restringir los indicadores sustantivos (tendencias, geografía, colaboración) a los registros principales (ismainarticle = True), salvo que se esté estudiando explícitamente la capa de referencias. Para corpora grandes, el enriquecimiento opcional con OpenAlex debe planificarse cuidadosamente: la resolución por DOI puede estar sujeta a límites de tasa, por lo que se recomiendan estrategias de *batching* y *caching* al escalar. Finalmente, los análisis posteriores deben apoyarse en las uniones relacionales definidas por el modelo de datos (SR, PersonID, journal_id), en lugar de reparsear cadenas crudas, asegurando que los resultados se mantengan consistentes, auditables y reproducibles entre ejecuciones.

6. CONCLUSIONES

Este estudio presentó BibFusion, una canalización de preprocesamiento reproducible que armoniza las exportaciones de Scopus y Web of Science en un único corpus trazable y listo para el análisis. BibFusion operacionaliza un modelo entidad-relación unificado y produce siete tablas relacionales —Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal y Scimagodb— respaldadas por claves primarias/foráneas explícitas y campos de procedencia que preservan la integridad relacional entre entidades. La canalización implementa normalización sistemática (estandarización ASCII/mayúsculas y depuración de SR/SR_ref), integración conservadora entre fuentes (desduplicación con prioridad al DOI y respaldos controlados), identificación canónica de autores mediante *PersonID* y consolidación estructurada de citaciones y outlets.

Un resultado central es la combinación de estructura relacional con auditabilidad: BibFusion genera no solo salidas estandarizadas para análisis bibliométricos y de redes, sino también artefactos de auditoría (bitácoras de fusión/enriquecimiento y archivos de alias/conflictos) que hacen transparentes las decisiones clave y permiten una revisión humana focalizada cuando los metadatos permanecen ambiguos. En conjunto, estos elementos proporcionan una base práctica para flujos de trabajo bibliométricos reproducibles, reduciendo inconsistencias evitables entre fuentes y documentando explícitamente la incertidumbre residual, en lugar de ocultarla.

BibFusion aporta valor práctico a la comunidad de cienciometría y bibliometría al transformar dos fuentes de datos heterogéneas y ampliamente utilizadas en un flujo de integración reproducible y auditable. Al estandarizar identificadores, consolidar duplicados y exponer una estructura entidad-relación clara, el paquete reduce la curaduría manual repetitiva y propensa a errores que suele preceder los análisis bibliométricos —en particular en estudios de colaboración y geografía científica, donde las inconsistencias de autores y afiliaciones pueden sesgar fuertemente los resultados—. De igual manera, BibFusion hace explícita la incertidumbre remanente mediante logs de auditoría y artefactos de conflicto y alias, lo que permite una inspección y corrección transparentes sin comprometer la reproducibilidad. Como resultado, los investigadores pueden invertir menos esfuerzo en limpieza ad hoc y más en análisis interpretables, manteniendo a la vez un rastro documentado que respalde la verificación, la extensión y la reutilización entre estudios y dominios de investigación.

Varias extensiones podrían fortalecer aún más BibFusion y ampliar su aplicabilidad. En primer lugar, el emparejamiento podría robustecerse incorporando señales adicionales de similitud más allá de la cascada conservadora actual —como concordancia estructurada en revista/volumen/páginas, umbrales calibrados de similitud de cadenas y modelos probabilísticos o aprendidos de *record linkage*—, manteniendo la auditabilidad y controlando las fusiones falsas. En segundo lugar, la cobertura podría ampliarse incorporando fuentes y formatos adicionales (p. ej., snapshots de OpenAlex, metadatos de Crossref, Dimensions, PubMed o repositorios institucionales) y mejorando la ingestión ante esquemas de exportación cambiantes. En tercer lugar, el aseguramiento de la

calidad podría reforzarse mediante tableros automatizados de control de calidad que reporten métricas de preparación (p. ej., completitud de DOI y país, tasas de consolidación de duplicados, resolubilidad del extremo citado y cobertura de identificadores de autor) y que destaquen clústeres de ambigüedad de alto impacto para su revisión. Por último, la escalabilidad del flujo puede mejorarse mediante caching y actualizaciones incrementales (p. ej., reejecutar solo el enriquecimiento o solo la fusión sobre registros recién añadidos), habilitando el mantenimiento eficiente de corpus longitudinales conforme las bases e identificadores evolucionan con el tiempo.

Agradecimientos

Los autores agradecen el apoyo de la Universidad Nacional de Colombia, MetricSci y el Data Lab de la Universidad Nacional de Colombia —Sede La Paz por facilitar el desarrollo y las pruebas de la canalización de preprocesamiento BibFusion, proporcionar alojamiento para el repositorio y equipos de cómputo, y apoyar a los estudiantes involucrados en este proyecto. Los autores también agradecen a OpenAI por proporcionar herramientas de inteligencia artificial (IA) utilizadas durante el desarrollo y la redacción (p. ej., Codex y ChatGPT) para apoyar la generación y optimización de partes del código, así como para respaldar la organización del manuscrito y su revisión iterativa. Todas las salidas producidas con apoyo de IA fueron revisadas y validadas por los autores, quienes asumen plena responsabilidad por el código y el manuscrito finales.

Financiación

Esta investigación recibió financiación de la Universidad Nacional de Colombia, Sede de La Paz, a través del proyecto “Estrategias Innovadoras de Marketing Emprendedor: Un Caso Aplicado al Laboratorio de Datos de la Universidad Nacional” (Hermes 64027).

Conflicto de intereses

Los autores declaran no tener conflictos de interés.

Declaración de contribuciones

Conceptualización; Software; Curaduría de datos; Metodología; Validación; Redacción – revisión y edición: Angelo Britto.

Validación; Investigación; Análisis formal; Visualización; Redacción – revisión y edición: Martha Zuluaga.

Conceptualización; Software; Metodología; Curaduría de datos; Análisis formal; Visualización; Redacción – borrador original; Redacción – revisión y edición; Administración del proyecto: Sebastian Robledo.

Declaración de consentimiento de datos

Los datos bibliográficos utilizados en este estudio se obtuvieron mediante exportaciones bajo licencia de Scopus y Web of Science y están sujetos a los términos y condiciones de sus respectivos proveedores. Por tanto, los archivos crudos exportados no se comparten públicamente. Las salidas procesadas generadas por BibFusion (tablas de entidades agregadas en CSV) y los detalles del código/flujo de trabajo pueden ponerse a disposición previa solicitud razonable, sujeto al cumplimiento de las restricciones de licencia de los proveedores de datos.

REFERENCIAS

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Chavarro, D., Alperin, J. P., & Willinsky, J. (2025). On the open road to universal indexing: OpenAlex and Open Journal Systems. *Quantitative Science Studies, 6,* 1039-1058. https://doi.org/10.1162/qss.a.17

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377. https://doi.org/10.1002/asi.20317

Chen, X., Mao, J., & Li, G. (2024). A co-citation approach to the analysis on the interaction between scientific and technological knowledge. *Journal of Informetrics, 18*(3), 101548. https://doi.org/10.1016/j.joi.2024.101548

Cioffi, A., Coppini, S., Massari, A., Moretti, A., Peroni, S., Santini, C., & Shahidzadeh Asadi, N. (2022). Identifying and correcting invalid citations due to DOI errors in Crossref data. *Scientometrics, 127*(6), 3593-3612. https://doi.org/10.1007/s11192-022-04367-w

Crystal-Ornelas, R., Varadharajan, C., O'Ryan, D., Beilsmith, K., Bond-Lamberty, B., Boye, K., Burrus, M., Cholia, S., Christianson, D. S., Crow, M., Damerow, J., Ely, K. S., Goldman, A. E., Heinz, S. L., Hendrix, V. C., Kakalia, Z., Mathes, K., O'Brien, F., Pennington, S. C., … Agarwal, D. A. (2022). Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. *Scientific Data, 9*(1), 700. https://doi.org/10.1038/s41597-022-01606-w

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130*(4), 2475-2492. https://doi.org/10.1007/s11192-025-05293-3

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies, 5*(1), 31-49. https://doi.org/10.1162/qss_a_00286

Delgado-Quirós, L., & Ortega, J. L. (2025). Citation counts and inclusion of references in seven free-access scholarly databases: A comparative analysis. *Journal of Informetrics, 19*(1), 101618. https://doi.org/10.1016/j.joi.2024.101618

Demaine, J. (2022). Fractionalization of research impact reveals global trends in university collaboration. *Scientometrics, 127*(5), 2235-2247. https://doi.org/10.1007/s11192-021-04246-w

Elstad, M., Ahmed, S., Røislien, J., & Douiri, A. (2023). Evaluation of the reported data linkage process and associated quality issues for linked routinely collected healthcare data in multimorbidity research: a systematic methodology review. *BMJ Open, 13*(5), e069212. https://doi.org/10.1136/bmjopen-2022-069212

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. *Journal of the Association for Information Science and Technology, 72*(8), 1039-1058. https://doi.org/10.1002/asi.24472

Kara, B. C., Şahin, A., & Dirsehan, T. (2025). BibexPy: Harmonizing the bibliometric symphony of Scopus and Web of Science. *SoftwareX, 30,* 102098. https://doi.org/10.1016/j.softx.2025.102098

Kim, J., & Owen-Smith, J. (2021). ORCID-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics, 126*(3), 2057-2083. https://doi.org/10.1007/s11192-020-03826-6

Kumpulainen, M., & Seppänen, M. (2022). Combining Web of Science and Scopus datasets in citation-based literature study. *Scientometrics, 127*(10), 5613-5631. https://doi.org/10.1007/s11192-022-04475-7

Lastilla, L., Ammirati, S., Firmani, D., Komodakis, N., Merialdo, P., & Scardapane, S. (2022). Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library. *Information Processing & Management, 59*(3), 102875. https://doi.org/10.1016/j.ipm.2022.102875

Lim, W. M., Kumar, S., & Donthu, N. (2024). How to combine and clean bibliometric data and use bibliometric tools synergistically: Guidelines using metaverse research. *Journal of Business Research, 182,* 114760. https://doi.org/10.1016/j.jbusres.2024.114760

Maisano, D. A., Mastrogiacomo, L., Ferrara, L., & Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics, 130,* 1901-1938. https://doi.org/10.1007/s11192-025-05244-y

Massari, A., Mariani, F., Heibi, I., Peroni, S., & Shotton, D. (2024). OpenCitations Meta. *Quantitative Science Studies, 5*(1), 50-75. https://doi.org/10.1162/qss_a_00292

Matveeva, N., Sterligov, I., & Lovakov, A. (2022). International scientific collaboration of post-Soviet countries: a bibliometric analysis. *Scientometrics, 127*(3), 1583-1607. https://doi.org/10.1007/s11192-022-04274-0

McKay, A. S. (2026). Common errors in bibliometric reviews and a novel method for correcting them. *Scientometrics*. https://doi.org/10.1007/s11192-026-05544-x

Mischo, W., Schlembach, M., & Cabada, E. (2024). Relationships between journal

publication, citation, and usage metrics within a Carnegie R1 university collection: A correlation analysis. *College and Research Libraries, 85*(2), 234-253. https://doi.org/10.5860/crl.85.2.234

Ng, J. Y., Liu, H., Masood, M., Syed, N., Stephen, D., Ayala, A. P., Sabé, M., Solmi, M., Waltman, L., Haustein, S., & Moher, D. (2025). Guidance for the reporting of bibliometric analyses: A scoping review. *Quantitative Science Studies, 6,* 988-1001. https://doi.org/10.1162/qss.a.12

Nikolić, D., Ivanović, D., & Ivanović, L. (2024). An open-source tool for merging data from multiple citation databases. *Scientometrics, 129*(7), 4573-4595. https://doi.org/10.1007/s11192-024-05076-2

Nowakowska, M. (2025). A comprehensive approach to preprocessing data for bibliometric analysis. *Scientometrics, 130*(9), 5191-5225. https://doi.org/10.1007/s11192-025-05415-x

Ornstein, J. T. (2025). Probabilistic record linkage using Pretrained text embeddings. Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association, *Advance online publication,* 1-12. https://doi.org/10.1017/pan.2025.10016

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In *arXiv [cs.DL].* https://doi.org/10.48550/ARXIV.2205.01833

Purnell, P. J. (2022). The prevalence and impact of university affiliation discrepancies between four bibliographic databases-Scopus, Web of Science, Dimensions, and Microsoft Academic. *Quantitative Science Studies, 3*(1), 99-121. https://doi.org/10.1162/qss_a_00175

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics, 15*(3), 101166. https://doi.org/10.1016/j.joi.2021.101166

Robledo, S., Valencia, L., Zuluaga, M., Echeverri, O. A., & Valencia, J. W. A. (2024). tosr: Create the Tree of Science from WoS and Scopus. *Journal of Scientometric Research, 13*(2), 459-465. https://doi.org/10.5530/jscires.13.2.36

Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX, 10*(100263), 100263. https://doi.org/10.1016/j.softx.2019.100263

Ruiz-Rosero, J., Ramirez-Gonzalez, G., & Viveros-Delgado, J. (2019). Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. *Scientometrics, 121*(2), 1165-1188. https://doi.org/10.1007/s11192-019-03213-w

Schmal, W. B. (2024). How transformative are transformative agreements? Evidence from Germany across disciplines. *Scientometrics, 129,* 1863-1889. https://doi.org/10.1007/s11192-024-04955-y

Sivertsen, G., Rousseau, R., & Zhang, L. (2025). The motivations for and effects of modified fractional counting. *Journal of Informetrics, 19*(3), 101681. https://doi.org/10.1016/j.joi.2025.101681

Vaccaro, G., Sánchez-Núñez, P., & Witt-Rodríguez, P. (2022). Bibliometrics evaluation of scientific journals and country research output of dental research in Latin America using Scimago Journal and Country Rank. *Publications, 10*(3), 26. https://doi.org/10.3390/publications10030026

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

Velez-Estevez, A., Perez, I. J., García-Sánchez, P., Moral-Munoz, J. A., & Cobo, M. J. (2023). New trends in bibliometric APIs: A comparative analysis. *Information Processing & Management, 60*(4), 103385. https://doi.org/10.1016/j.ipm.2023.103385

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies, 2*(1), 20-41. https://doi.org/10.1162/qss_a_00112

Wang, F., Dong, J., Lu, W., & Xu, S. (2023). Collaboration prediction based on multilayer all-author tripartite citation networks: A case study of gene editing. *Journal of Informetrics, 17*(1), 101374. https://doi.org/10.1016/j.joi.2022.101374

Yang, J., Wu, L., & Lyu, L. (2024). Research on scientific knowledge evolution patterns based on ego-centered fine-granularity citation network. *Information Processing & Management, 61*(4), 103766. https://doi.org/10.1016/j.ipm.2024.103766

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: possible reasons, implications, and solutions. *Scientometrics, 129,* 5869-5891. https://doi.org/10.1007/s11192-023-04923-y

APÉNDICES

Apéndice A: Diagramas de flujo específicos por fuente

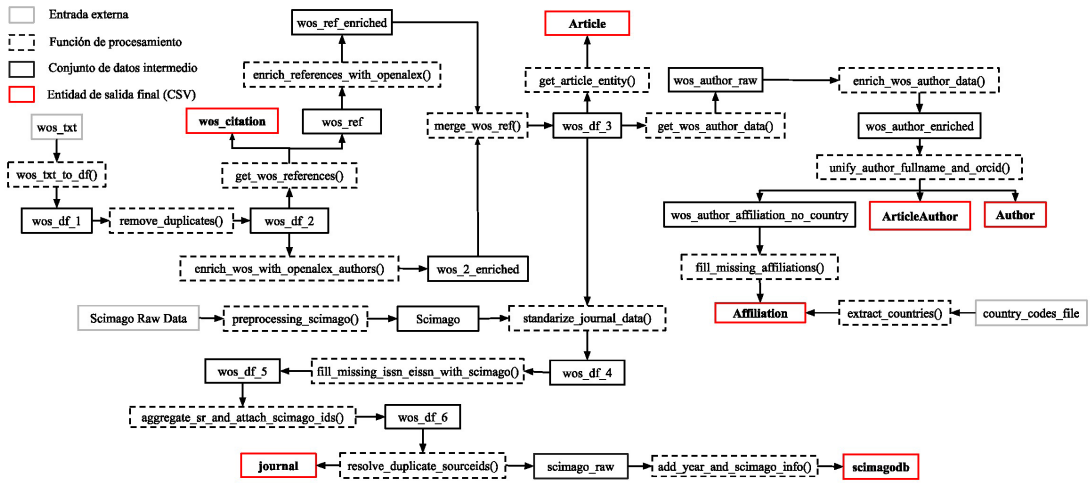

Figura A1. Flujo de trabajo a nivel de función para el preprocesamiento de WoS en TXT. **Nota:** Este diagrama ofrece una vista a nivel de función de las transformaciones aplicadas a la exportación en texto plano de WoS, complementando el flujo de extremo a extremo de la Figura 1. Traza la secuencia desde la ingestión y canonicalización hasta la normalización de metadatos (p. ej., estandarización de títulos y de claves SR), el parseo de afiliaciones con extracción de país y la generación de salidas estandarizadas de staging utilizadas en la fusión posterior entre bases. Los recuadros delineados en rojo indican las salidas estandarizadas de entidades que corresponden al modelo ER unificado de la Figura 2.

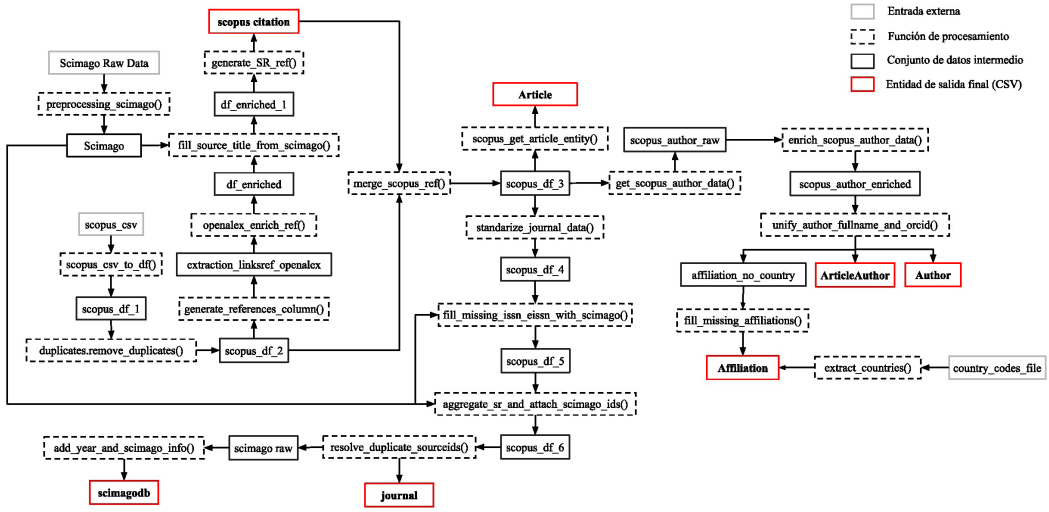

Figura A2. Flujo de trabajo a nivel de función para el preprocesamiento de Scopus en CSV y el enriquecimiento en OpenAlex basado en DOI. Nota: Este diagrama complementa la Figura 1 al detallar la secuencia de funciones aplicadas a las exportaciones CSV de Scopus. Muestra la ingestión y la

normalización de campos (incluida la estandarización de DOI), el enriquecimiento impulsado por DOI vía OpenAlex (p. ej., identificadores de obra e identificadores de autor cuando están disponibles) y pasos posteriores como la construcción de referencias/SR, la consolidación de revistas/Scimago y el parseo de afiliaciones con extracción de país. El flujo genera entidades de staging estandarizadas y alineadas por fuente que alimentan la etapa posterior de desduplicación y fusión entre bases.

Apéndice B: Configuraciones de exportación utilizadas para la recolección de datos (Scopus y Web of Science)

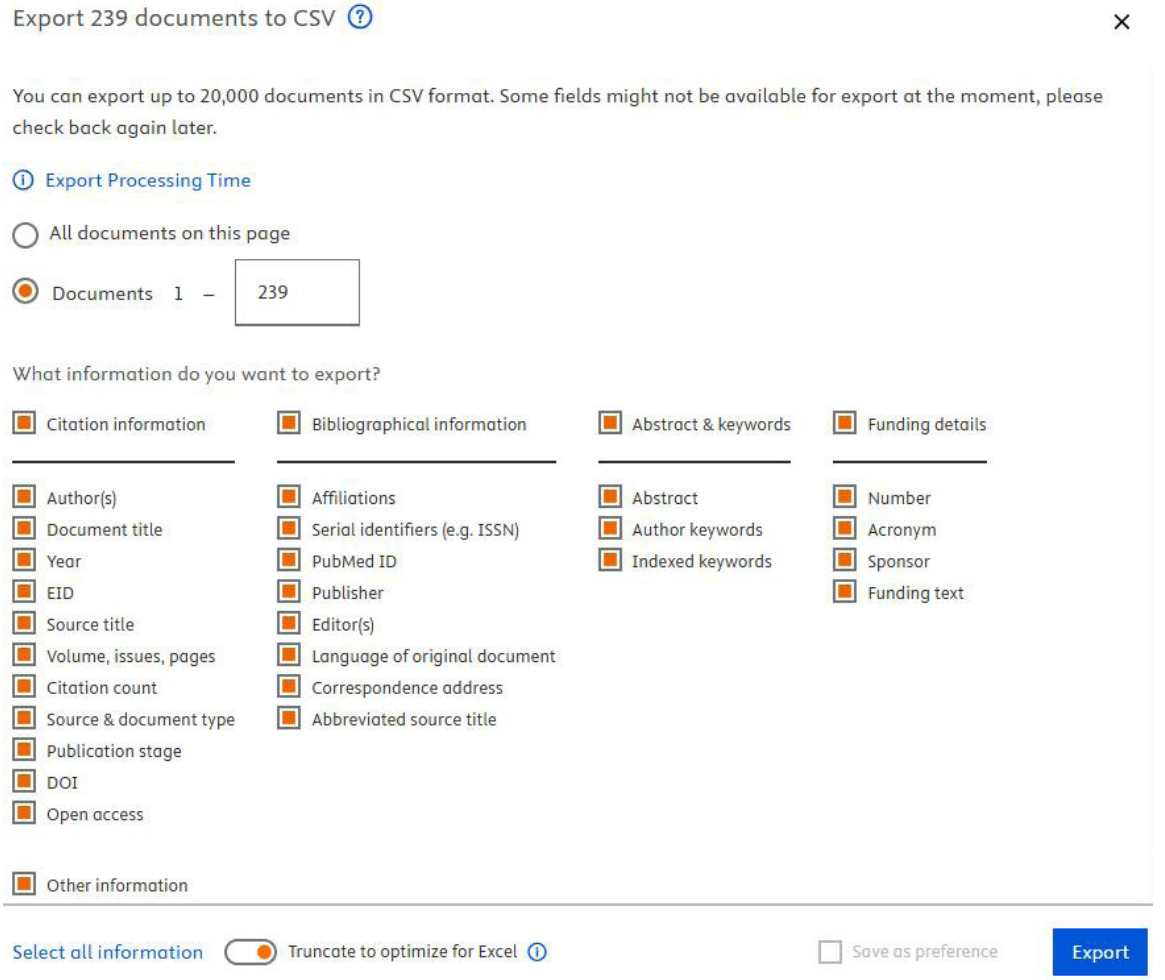

Figura B1. Configuración de exportación CSV de Scopus ("Select all information"). **Nota:** Esta figura documenta las configuraciones de exportación de Scopus utilizadas para generar el insumo crudo en CSV para BibFusion, en las cuales la exportación se configura para incluir toda la información disponible (p. ej., información de citaciones, información bibliográfica, resúmenes/palabras clave, detalles de financiación y otros descriptores). Estas configuraciones son necesarias para preservar metadatos completos y asegurar una normalización, desduplicación y trazabilidad consistentes en etapas posteriores.

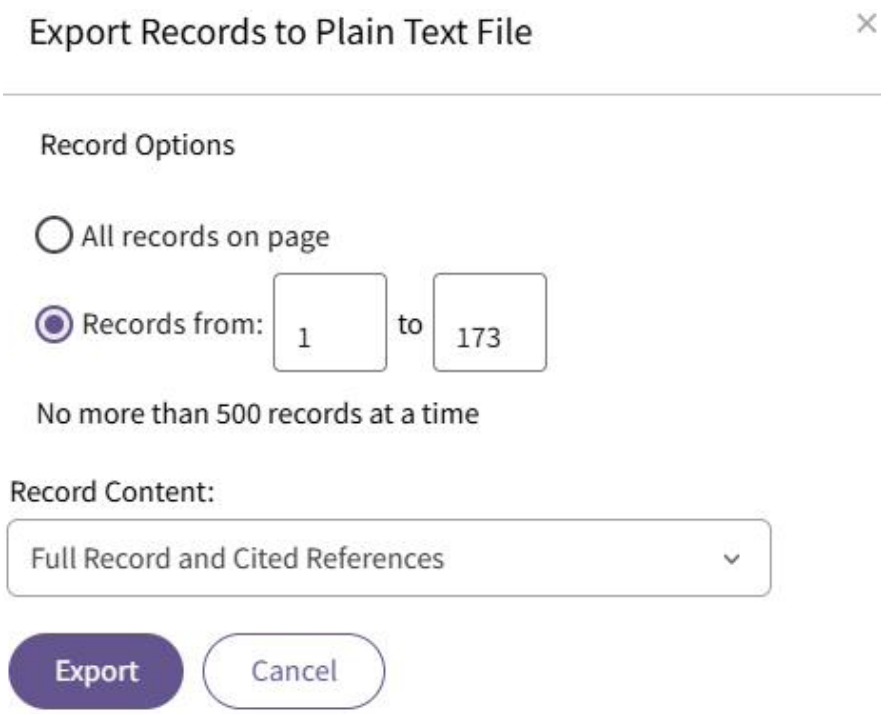

Figura B2. Configuración de exportación TXT de Web of Science ("Full Record and Cited References"). **Nota:** Esta figura documenta las configuraciones de exportación de Web of Science Core Collection utilizadas para generar el insumo crudo en TXT para BibFusion, donde los registros se exportan como archivo de texto plano con "Full Record and Cited References" seleccionado. Esta configuración garantiza que tanto los metadatos bibliográficos completos como las cadenas de referencias citadas estén disponibles para la normalización SR/SR_ref y la construcción de enlaces de citación en el corpus unificad.