Vol. 6, No. 1, 2026, 1-21. DOI: 10.47909/ijsmc.342

pro-metrics

ORIGINAL ARTICLE

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

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.

Received: 29-11-2025. **Accepted:** :09-02-2026. **Published:** 15-02-2026.

How to cite: 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), 1-21. DOI: 10.47909/ijsmc.324

1. INTRODUCTION

Bibliometric analyses are widely used to trace how research domains evolve; however, the validity of these insights depends on the consistency of the underlying bibliographic metadata (Zhang *et al.,* 2024). In practice, integrating records from major sources such as Scopus and Web of Science (WoS) remains challenging because the two platforms differ in coverage, export structures, and field completeness (Kumpulainen & Seppänen, 2022). These discrepancies frequently result in the replication of items across sources, variations in author names, ambiguous or incomplete affiliations, and non-standardized references that fragment citation links. Consequently, downstream indicators —including country and institutional maps, temporal production trends, and coauthorship or citation networks— may exhibit bias or instability if preprocessing is executed in an ad hoc manner. A reproducible harmonization step prior to any domain-specific analysis (e.g., entrepreneurial marketing-related searches) is therefore essential to build a clean, integrated corpus suitable for reliable scientometric analysis (Nowakowska, 2025).

Integrating Scopus and Web of Science into a single corpus is difficult because the two databases differ in coverage, export structure, and metadata completeness (Delgado-Quirós & Ortega, 2024). DOIs may be missing or malformed; textual fields (titles and cited-reference strings) vary in capitalization, punctuation, and character normalization; and author and affiliation data are often heterogeneous or incomplete, complicating attribution at the person and country levels (Visser *et al.,* 2021). Taken together, these discrepancies inflate duplicates, introduce ambiguity in author identities, and fragment citation links, thereby biasing indicators of productivity, collaboration, and scientific geography (Nowakowska, 2025).

To address this problem, we present BibFusion, a reproducible and auditable software tool that harmonizes and merges Scopus and Web of Science exports into a unified, traceable, analysis-ready dataset. BibFusion standardizes metadata, parses affiliations (including country extraction), and optionally enriches records through DOI-based resolution against OpenAlex to recover persistent work and author identifiers when available (Nowakowska, 2025; Priem *et al.,* 2022). Integration follows a conservative deduplication strategy (DOI-first, with fallback rules when missing) and consolidates authorship to support consistent author-level analyses (Visser *et al.,* 2021). The result is a relational corpus materialized as linked tables, accompanied by audit artifacts that document provenance and merge decisions to facilitate transparent review and reproducibility (Maisano *et al.,* 2025).

The study's scope is delineated by the search query implemented in Scopus and WoS, with a focus on English-language journal articles within the stipulated publication window for the primary corpus. Given that cross-source integration is most reliable when persistent identifiers are available, BibFusion prioritizes DOI-based matching when present and relies on conservative metadata agreement rules when DOIs are missing or malformed. Similarly, the processes of enrichment and author consolidation are enhanced by the utilization of external identifiers, such as ORCID or OpenAlexAuthorID. However, these identifiers may exhibit incompleteness across different sources. Consequently, unresolved cases are retained without the imposition of forced decisions, and they are exposed through audit outputs for transparent review. It is important to note that while the main corpus is query-bounded, cited references may extend beyond the query's time window and document-type constraints. This is expected when constructing citation networks.

This study proposes a reproducible and audit-ready preprocessing pipeline that harmonizes Scopus and WoS exports into a unified relational corpus. The workflow integrates identifier-centric enrichment (via DOI resolution in OpenAlex) with systematic author disambiguation using a canonical PersonID. It materializes the merged corpus as linked entities with explicit provenance. In addition to analysis-ready outputs for productivity, collaboration, and geographic analyses, BibFusion generates audit artifacts (e.g., aliases, potential conflicts, and merge logs) that facilitate transparency and reproducibility in integration decisions. The remainder of this study is structured as follows. Section 1.1 reviews related work on cross-database bibliographic integration and disambiguation. Section 2 delineates the

composition of the package, including its inputs and outputs, as well as the unified data model. Section 3 delineates the preprocessing methodology, encompassing normalization, OpenAlex-based enrichment, deduplication and merge rules, author disambiguation, citation cleaning, and journal/Scimago consolidation. Section 4 reports the resulting corpus and key quality control indicators. Section 5 delves into the implications, limitations, and practical guidance for reuse. Section 6 concludes and outlines future improvements. Supplementary workflow details and documentation can be found in the Appendices.

1.1. Literature review

Existing bibliometric ecosystems provide robust support for importing, mapping, and visualizing Scopus and WoS exports, including widely used packages and interfaces such as bibliometrix/biblioshiny (Aria & Cuccurullo, 2017; Maisano *et al.,* 2025), VOSviewer (van Eck & Waltman, 2010), and CiteSpace (Chen, 2006), as well as Python-oriented tools for access and trend analysis such as pybliometrics (Rose & Kitchin, 2019) and ScientoPy (Ruiz-Rosero *et al.,* 2019). While these tools are highly effective once data are ingested, cross-database integration typically still requires additional processing to reconcile heterogeneous export schemas and metadata inconsistencies (e.g., missing or inconsistently recorded DOIs, variant author strings, and incomplete or non-standardized affiliations), particularly when the goal is to preserve and harmonize cited references across sources. In response to these challenges, integration-focused workflows have emerged. These range from tosr, which operationalized Scopus-WoS integration while explicitly incorporating reference information beyond standard merges (Robledo *et al.,* 2024), to newer Python pipelines such as BibexPy, which emphasizes automated merging and metadata enhancement via external services (Kara *et al.,* 2025), and configurable matching toolkits such as TeslaSCItoolkit, which formalize record linkage through similarity metrics and match classification (Nikolić *et al.,* 2024). A review of the extant literature reveals a general tendency toward identifier-centric and automation-oriented integration. However, there are notable lacunae with respect to (i) the production of an explicit, relational entity-relationship (ER) representation that differentiates core entities from linking tables, and (ii) the provision of auditable, versioned artifacts that document merge and disambiguation decisions, particularly at the reference level.

As demonstrated in the extant literature, structural challenges have been thoroughly documented in the context of bibliometric integration. These challenges persist in impeding the reliable consolidation of Scopus and WoS records (McKay, 2026). First, the efficacy of deduplication is often undermined by missing, malformed, or inconsistently recorded DOIs. This forces reliance on secondary matching rules (e.g., title-year-first author) and increases both false positives and false negatives (Culbert *et al.,* 2025). Second, the process of author disambiguation remains imperfect due to orthographic variation, inconsistent use of initials, transliteration differences (accented forms vs. ASCII), and homonymy. These factors can lead to the fragmentation or conflation of identities, resulting in the distortion of productivity and collaboration indicators (Kim & Owen-Smith, 2021). Third, cited-reference information is characterized by significant heterogeneity. Reference strings frequently exhibit a variety of formats, irregular punctuation, standalone years, or malformed entries. These characteristics can disrupt citation links and weaken citation and co-citation networks, particularly when reference-level integration is necessary across sources (Cioffi *et al.,* 2022). In conclusion, outlet and affiliation metadata frequently necessitate additional normalization (e.g., journal name/abbreviation harmonization, incomplete affiliations, and missing or implicit country information), which can introduce bias in collaboration and scientific geography measures (Purnell, 2022). The aforementioned challenges, when considered collectively, serve as a compelling rationale for the development of reproducible preprocessing pipelines. These pipelines are designed to standardize metadata, enhance deduplication, improve traceability, and optimize citation links prior to bibliometric and network analysis.

Despite the sustained progress in bibliometric software and database-specific importers, an important gap persists in cross-database

integration. Many workflows emphasize downstream indicators and visualization, while the integration layer (including cited-reference harmonization) remains difficult to reproduce and audit (Ng *et al.,* 2025). In particular, relatively few approaches formalize the merged output as an explicit ER model that separates core entities (e.g., Articles, Authors, Journal, Affiliation) from linking tables (e.g., ArticleAuthor, Citation) and maintains stable identifiers across sources (Massari *et al.,* 2024). Consequently, pivotal steps —such as deduplication decisions, author identity resolution, and citation-string cleaning— are frequently implemented in ways that are challenging to inspect, validate, or rerun consistently when inputs change. Furthermore, the production of audit artifacts (e.g., aliases, potential conflicts, and merge logs) is not uniform, which necessitates the reliance of researchers on manual spot checks with limited documentation (Elstad *et al.,* 2023). This approach fosters the establishment of a reproducible pipeline, which culminates in the generation of an analysis-ready relational corpus, accompanied by auditable files. These files serve to enhance the transparency of integration decisions, thereby facilitating systematic verification and iterative refinement.

In this context, BibFusion is positioned as a reproducible, audit-ready integration layer between Scopus and WoS that complements —rather than replaces— existing bibliometric analysis ecosystems. The integration gap is addressed by materializing the merged output as an explicit ER structure and by preserving reference-level links through dedicated citation and linking tables (e.g., standardized SR/SR_ref keys and a cleaned citation edge list; Delgado-Quirós & Ortega, 2025). Methodologically, BibFusion employs a DOI-first matching strategy with conservative fallbacks. When DOIs can be resolved, BibFusion enriches records via OpenAlex to recover persistent identifiers for works and authors (Chavarro *et al.,* 2025). It also assigns a canonical PersonID (ORCID → OpenAlexAuthorIDx → normalized name) to strengthen author-level continuity, where identifier coverage supports reliable linking (Rehs, 2021). Of particular significance is the externalization of uncertainty through audit outputs (aliases, potential conflicts, and merge logs), thereby enabling transparent inspection and iterative correction (Ornstein, 2025). The result is a unified, analysis-ready relational corpus with explicit provenance that can be directly consumed for bibliometric, collaboration, geographic, and citation-network analyses.

2. PACKAGE OVERVIEW AND DATA MODEL

2.1. Package purpose, inputs, and outputs

BibFusion is a modular preprocessing package that harmonizes bibliographic exports from Scopus and WoS into a unified, traceable, analysis-ready corpus. The system has been developed to ingest raw Scopus CSV and WoS TXT files, apply systematic normalization to key metadata fields (e.g., titles, reference keys, affiliations, and country information), and enrich records through OpenAlex to recover persistent identifiers for works and authors when available (Velez-Estevez *et al.,* 2023). The pipeline then performs DOI-first deduplication with conservative fallbacks (Kara *et al.,* 2025), assigns canonical PersonIDs for author disambiguation, cleans and remaps citation strings to preserve consistent citation links, and consolidates journal and Scimago-related information using ISSN/EISSN rules (Mischo *et al.,* 2024). The workflow has been implemented as a series of reusable modules, including ingestion, normalization, enrichment, matching/merging, author disambiguation, and citation/journal processing. This modular approach enables users to execute the workflow in its entirety to generate a complete dataset or to run individual stages to support incremental updates and targeted quality control.

2.2. Workflow at a glance

Figure 1 provides a synopsis of the end-to-end workflow that has been implemented by BibFusion. Scopus CSV and WoS TXT exports are processed in parallel through a series of modular stages. These stages include ingestion and canonicalization, metadata normalization (including affiliation parsing and country extraction), and optional DOI-based enrichment via OpenAlex (Culbert *et al.,* 2025). The standardized records are then deduplicated and merged using a DOI-first matching cascade with conservative fallbacks. The integrated

corpus is subsequently structured through PersonID-based author disambiguation, citation-string cleaning and remapping (to populate the Citation table), and journal/Scimago consolidation using ISSN/EISSN rules. The outputs are materialized as seven relational entity tables and accompanied by audit artifacts that document aliases, potential conflicts, and merge logs. Function-level process diagrams (including intermediate data frames and transformation functions) are provided in Appendix A (Figures A1 and A2).

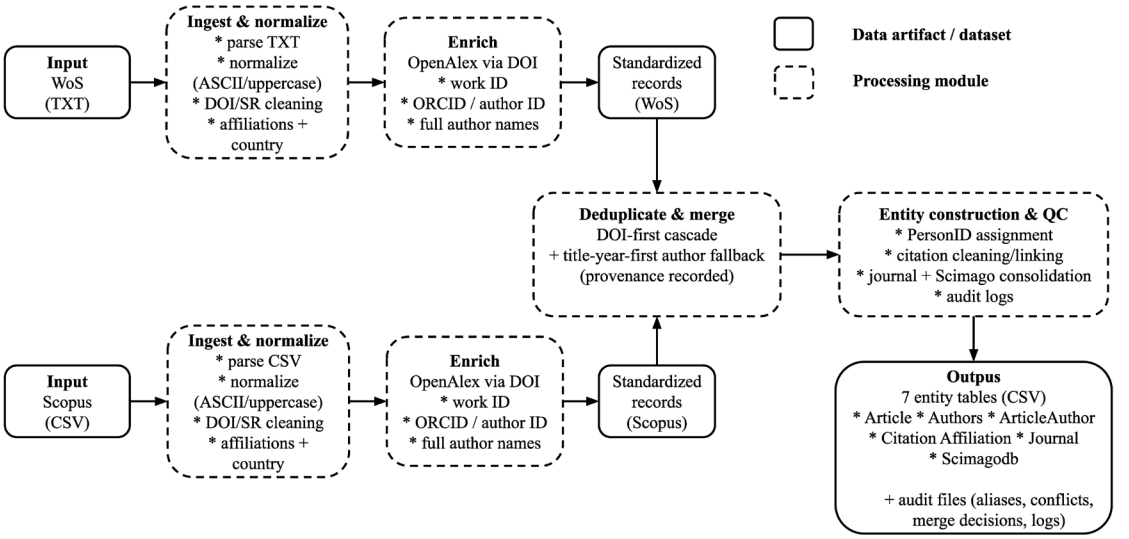

Figure 1. Overview of the preprocessing pipeline. **Note:** Prepared by the authors.

2.3. Data model

Figure 2 presents the unified ER model of the integrated corpus produced by BibFusion. The pipeline first maps WoS and Scopus exports into source-specific staging tables that share the same entity structure and key conventions (a shared, aggregated staging schema) after parsing and normalization. At this stage, standardizing structure is preferable to merging heterogeneous exports directly. This approach reduces discrepancies in field naming and formatting and facilitates consistent, rule-based deduplication and integration (Rehs, 2021). The final corpus is stored as seven linked relational entities: Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal, and Scimagodb. As illustrated in Figure 2, solid connectors denote enforced joins, while dashed connectors signify FK-like or optional resolutions when coverage is incomplete (e.g., SR_ref in Citation may not always resolve to a full Articles record). Provenance is preserved through source indicators (e.g., sources_merged) and main-vs-reference flags (e.g., ismainarticle), thereby enabling reproducible bibliometric, collaboration, geographic, and citation-network analyses with transparent filtering of record types (Lastilla *et al.*, 2022).

2.4. Output inventory

Table 1 enumerates the seven CSV outputs produced by BibFusion and their roles in the unified relational schema. The following categories are included in the database: Articles (work-level records), Authors (canonical identities keyed by PersonID), ArticleAuthor (work-author links), Citation (directed citation edges), Affiliation (work-author-affiliation instances with country), Journal (normalized outlets keyed by journal_id), and Scimagodb (journal-level metrics joinable via journal_id). According to the findings reported in Table 1, the row granularity, primary key(s), and join fields that preserve relational integrity for each file are documented. The join fields include SR, PersonID, journal_id, and SR → SR_ref links. It should be noted that SR_ref may not always resolve to a full Articles record. In summary,

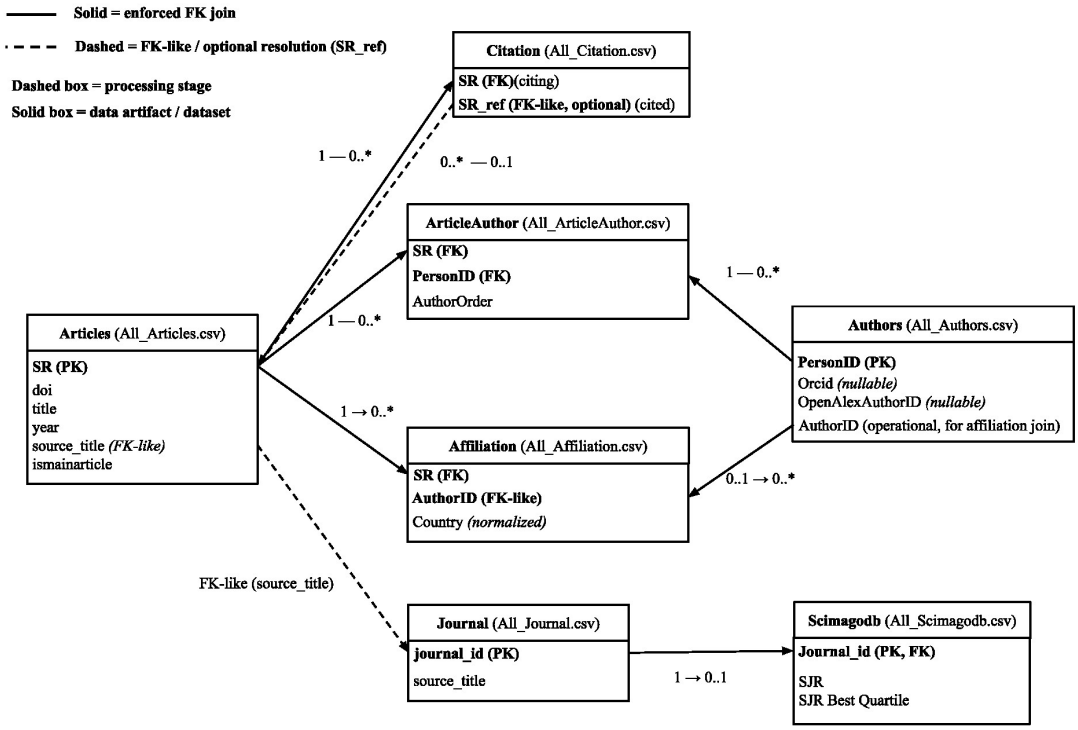

Figure 2. Unified ER model of the BibFusion integrated corpus. **Note:** Prepared by the authors.

the table offers a succinct reference manual for the assembly of reproducible trend, collaboration, geographic, and citation-network analyses from the standardized entities.

In addition to the entity tables summarized in Table 1, BibFusion organizes outputs into three directories. These directories separate source-level intermediate artifacts from the final integrated corpus. WoS_results/ and Scopus_results/ are where standardized staging outputs and source-specific logs are stored. These logs are generated during ingestion, normalization, and (when enabled) OpenAlex enrichment. This enables partial reruns and targeted debugging at the source level. The final deliverables, prepared for analysis, are stored in all_data_wos_scopus/ directory. This directory contains the seven relational CSV entities, along with audit files that document aliases, potential conflicts, and merge logs. This separation supports reproducible execution while ensuring that intermediate processing products remain distinct from the consolidated dataset that is subsequently utilized in downstream analyses. For transparent reuse, a comprehensive versioned data dictionary (variable definitions, provenance, and transformation rules) is distributed with the BibFusion repository (https://pypi.org/project/bibfusion).

3. METHODOLOGY

3.1. Data collection, exports and search strategy

BibFusion ingests two primary bibliographic sources: Scopus exports in CSV format and WoS TXT format. The records are obtained through each platform's standard export interface and include core metadata (e.g., title, authors, affiliations, publication year, source/journal information, DOI when available, cited references, and citation counts). The utilization of both sources serves to expand the scope of coverage while maintaining the provenance of the sources for the purpose of traceability. This enables the pipeline to harmonize records that refer to the same publication across various databases. Operationally, BibFusion employs source-specific parsers and maps both exports into an identical staging schema (source-aligned tables). Consequently, subsequent normalization, enrichment, and deduplication/merge steps

Output CSV	Row granularity	Primary key (PK)	Foreign keys / links (how to join)	Key columns (examples)	Purpose
All_Articles.csv	One row per *work record* (main records + reference-like records) after harmonization/merge	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	Canonical works table for trend/productivity analyses and as the node list for collaboration and citation networks
All_Authors.csv	One row per *unique person identity* after disambiguation	PersonID	PersonID is referenced by *All_ArticleAuthor*	PersonID, AuthorFullName, AuthorName, Orcid, OpenAlexAuthorID, ResearcherID, email, AuthorID	Canonical author identity table enabling author-level analyses and consistent joins across outputs
All_ArticleAuthor.csv	One row per *authorship link* (person-work association)	SR, PersonID, AuthorOrder	SR → *All_Articles.SR;* PersonID → *All_Authors.PersonID*	SR, PersonID, AuthorOrder, CorrespondingAuthor, OpenAlexAuthorID, AuthorID, openalex_work_id	Implements the many-to-many authorship relationship (coauthorship networks, author-order analyses, productivity by author)
All_Citation.csv	One row per *directed citation edge* (citing → cited)	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	Edge list for citation/co-citation networks and cleaned reference-linking across sources
All_Affiliation.csv	One row per *(work-author-affiliation) instance*	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	Institutional and geographic mapping (affiliations/countries), supporting collaboration-by-country and affiliation-based analyses
All_Journal.csv	One row per *normalized journal/source*	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	Normalized journal metadata enabling outlet-based analyses and a bridge to Scimago-style metrics
All_Scimagodb.csv	One row per *journal with available metrics*	journal_id	journal_id → *All_Journal.journal_id*	journal_id, title, Issn, eIssn, SJR, SJR best quartile, H index, publisher, country, categories, areas, coverage	Journal-level contextual metrics (quartiles/SJR/H-index, categories) for outlet benchmarking and stratified analyses

Table 1. BibFusion output inventory and relational join keys. Note: SR_ref is a normalized cited-reference key and does not always resolve to a full record in All_Articles (e.g., books, reports, or out-of-coverage works). In those cases, the citation edge is retained for citation-network analyses using the SR → SR_ref structure. PersonID is the canonical author identifier used across outputs, assigned using a priority rule (ORCID → OpenAlexAuthorID → normalized name) and can be populated via enrichment even when absent in the raw exports. Data generated by BibFusion.

operate on harmonized representations rather than heterogeneous raw formats. It is noteworthy that all exports were generated under the "full record" settings, inclusive of the cited references, as detailed in Appendix B. The records were retrieved using aligned selection criteria in Scopus and WoS to maximize the comparability of the sources. In both databases, an exact phrase search was conducted in the Title field for the term "entrepreneurial marketing" to prioritize precision and reduce topical noise relative to abstract- or keyword-based searches. The results of the study were restricted to English-language journal articles (document type: Article) published within the timeframe of 2005-2025 (inclusive). The Scopus query was implemented using the advanced search syntax (operationalized as PUBYEAR > 2004 AND PUBYEAR < 2026, with DOCTYPE "ar" corresponding to Article), whereas the WoS query was implemented using Core Collection field tags (e.g., TI for Title and PY for publication year). In the WoS database, the PY field may indicate either the final publication year or an early-access year, depending on the context. BibFusion is designed to preserve source-specific year fields while standardizing integrated outputs to ensure consistent downstream reporting.

Scopus (advanced search):

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

Web of Science Core Collection (advanced search):

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

BibFusion has been developed to function with full-record exports from both Scopus and WoS. Accordingly, the workflow assumes that (i) Scopus records are exported as CSV with "Select all information" enabled and (ii) WoS records are exported as TXT using "Full Record and Cited References." These settings ensure the availability of the minimum information required for harmonization and integration, including bibliographic descriptors (e.g., title, year, document type, source/journal identifiers such as ISSN/EISSN, and pagination where available), persistent identifiers (especially DOI), authorship and affiliation/address fields (for authorship linking and country extraction), and complete cited-reference strings (to construct and normalize SR/SR_ref keys and populate the Citation entity). To prevent information loss, any truncation option intended for spreadsheet compatibility (e.g., "optimize for Excel") should be disabled, as it has the potential to shorten long fields such as affiliations, abstracts, and cited references and to compromise matching and citation-link construction. The export configurations utilized in this study are documented in Appendix B (Figures B1 and B2). BibFusion relies on exports that collectively cover five metadata components. (i) The following bibliographic core metadata elements are required for normalization and consolidation of outlets: title, publication year, and source/journal information (including ISSN/EISSN when available), together with volume/issue/pages. (ii) Persistent and database-specific identifiers, most notably the DOI, serve as the primary anchors for cross-source matching and deduplication. (iii) Authorship and affiliation/address fields enable authorship linking, institutional parsing, and country extraction for geographic analyses. (iv) Full cited-reference strings populate the Citation entity. This, in turn, supports SR/SR_ref normalization. The purpose of this support is to preserve consistent citation links. (v) The retention of supplementary descriptors (e.g., citation counts, open-access indicators, and keywords/abstracts) is essential for maintaining traceability and facilitating downstream descriptive and thematic analyses following the generation of a unified corpus.

3.2. Data quality considerations and limitations

Given that BibFusion necessitates complete exports from Scopus and WoS, the pipeline has been engineered to address real-world missingness and inconsistencies that persist even under comprehensive export configurations.

Common issues include (i) empty or partially populated metadata fields (especially affiliations, addresses, and identifiers); (ii) missing, malformed, or inconsistently formatted DOIs, which can prevent exact matching across sources; and (iii) heterogeneity in author and source strings, including variations in punctuation, ordering, initials, and diacritics (accented forms vs. ASCII). BibFusion addresses these issues through systematic normalization (ASCII/uppercase standardization for key text fields such as titles and SR keys), DOI cleaning/normalization when DOIs are present, and a conservative matching cascade that prioritizes exact DOI alignment and applies a light title-year-first author fallback only when the DOI is unavailable. Rather than discarding records with incomplete fields, the pipeline retains them whenever a canonical work/reference key can be constructed. This approach preserves provenance while reducing the risk of false merges. The consolidation of author identities is achieved through a canonical PersonID (ORCID → OpenAlexAuthorID → normalized name), and the cleansing of cited-reference strings to remove malformed or low-information entries (e.g., standalone years) that would otherwise disrupt the integrity of citation linking. Residual ambiguities are not resolved without explicit articulation: BibFusion generates audit artifacts (i.e., aliases, potential conflicts, and merge logs) to facilitate the review and correction of uncertain cases without compromising reproducibility.

3.3. Corpus definition

The fundamental unit of analysis in BibFusion is the bibliographic record, as it is represented within the integrated corpus. Two record types are distinguished from the outset: main records, which are publications retrieved directly from Scopus and WoS under the specified query and filters, and reference records, which are cited items parsed from the reference lists of the main records. It is imperative to note that both entities are stored within the Articles entity, with the objective of maintaining a singular node space for the purpose of network construction. However, it is crucial to acknowledge a substantial discrepancy between them with respect to completeness. Typically, main records are characterized by their richness in metadata and the presence of persistent identifiers (e.g., DOI and source fields), while reference records may contain a cleaned/normalized reference key (SR/SR_ref) and a restricted set of attributes solely. Directed citation relations are represented in the Citation entity as edges from a citing main record (SR) to a cited reference key (SR_ref). It should be noted that not all cited items necessarily map to a fully described article record within the dataset. The Authorship relations are captured in ArticleAuthor exclusively when sufficient metadata is available for a reliable linkage to a canonical PersonID. This representation facilitates analyses centered on the retrieved corpus (e.g., productivity and geographic indicators on main records) while enabling citation-network analyses that transparently incorporate reference-only nodes without overstating their metadata quality.

The integrated Articles table, which was generated using the search strategy outlined in Section 3.1, contains a total of 8,569 bibliographic records. Of these, 253 are main records retrieved directly from Scopus and WoS (ismainarticle = True), while 8,316 are reference records parsed from cited-reference strings (ismainarticle = False). Provenance is retained in the merged sources, with 5,178 records observed in both the Scopus and WoS databases. Of these, 2,541 are Scopus-only records, while 850 are WoS-only records. A total of 183 records were found to be matched across both sources, 59 of which were exclusively found in Scopus, and 11 were exclusively found in WoS. The overall DOI availability rate is high (8,054 out of 8,569 records, representing 94% of the total) and remains substantial for main records (220 out of 253 records, representing 87% of the total). Notably, the DOI coverage among main records in both sources reached 98.9%, reflecting the complementarity of the two databases for identifier recovery. The primary corpus encompasses the period from 2005 to 2025, as delineated by the query. However, reference records extend beyond this timeframe, reaching as far back as 1900 and extending to 2026. This is due to the inheritance of years from cited items and the potential inclusion of infrequent early-access or noisy year values.

3.4. BibFusion pipeline and reproducibility

The workflow, as depicted in Figure 1, serves as a foundational framework for the subsequent discussion. Building upon this foundation, BibFusion implements a modular, end-to-end data processing pipeline that transforms raw Scopus CSV and WoS TXT exports into a harmonized, deduplicated, and auditable relational corpus. This corpus is designed to be suitable for reproducible bibliometric and network analyses. BibFusion's initialization process entails the ingestion of Scopus CSV and WoS TXT exports, which are parsed using source-specific parsers. These parsers facilitate the conversion of heterogeneous raw fields into a unified staging schema, while preserving source provenance to ensure traceability. Subsequently, a uniform normalization layer is applied to stabilize downstream matching. Key text fields (e.g., titles and author strings) are converted to a consistent ASCII/uppercase representation and trimmed to remove spurious whitespace and formatting artifacts. Concurrently, reference keys undergo standardization through a process of cleaning and normalization of SR/SR_ref strings. This involves the removal of inconsistent punctuation and low-information tokens, thereby ensuring the integrity and reliability of citation linking and deduplication. In conclusion, the affiliation and address fields are parsed to derive a normalized country attribute. Repetitive or inconsistent country mentions within a record are consolidated so that country-level and geographic indicators remain comparable across Scopus and WoS after integration.

Subsequent to normalization, BibFusion has the capacity to optionally enrich records using DOI-based resolution against OpenAlex. The purpose of this process is to recover persistent identifiers and improve cross-source linking. When a valid DOI is available, the pipeline queries OpenAlex to obtain a stable work identifier (e.g., OpenAlex work ID) and to expand author metadata, including OpenAlexAuthorID and ORCID, where available. This process later strengthens author disambiguation and joins across entities. This enrichment step is intentionally conservative and non-blocking: records with missing or invalid DOIs, or DOIs that cannot be resolved, are skipped and logged rather than causing failures, so that incomplete records remain in the integrated corpus while unresolved identifiers and enrichment outcomes are transparently documented for review. BibFusion integrates Scopus and WoS records through a DOI-first deduplication and merge strategy, applying exact matching on normalized DOIs whenever available and using a conservative fallback based on title-year-first author agreement when DOI is missing. In the event of the consolidation of two records, the pipeline employs a transparent best-record policy, which entails the retention of the most informative values across fields that are present in more than one record. This process also preserves provenance through the use of sources_merged and related flags, which indicate the origin of each integrated row. The unified corpus is then materialized into the relational entities defined in Section 2: authors are disambiguated using a canonical PersonID (ORCID → OpenAlexAuthorID → normalized name) and propagated to Authors and ArticleAuthor; citation strings are cleaned and standardized into directed edges in Citation; and outlet metadata are consolidated into Journal and linked to Scimagodb using ISSN/EISSN- and title-based normalization rules.

BibFusion integrates quality control measures at various stages of the workflow to minimize noise without compromising uncertainty. During the process of citation cleaning, low-information or malformed reference entries are filtered. Examples of such entries include empty SR/SR_ref keys, standalone years, or non-informative punctuation strings. Potentially ambiguous situations, such as borderline deduplication matches, author name collisions, or incomplete identifier resolution, are logged and surfaced. Rather than being forced into a single interpretation, these situations are instead logged and surfaced to allow for the possibility of multiple interpretations. In addition to the seven entity tables, the pipeline generates audit artifacts (e.g., aliases, potential conflicts, and merge logs) that provide an explicit trail of key decisions and cases requiring inspection. BibFusion facilitates reproducible diagnostics by providing a summary of core QC indicators, including duplicates removed under the matching cascade, DOI coverage (both overall and for main records), the completeness of country

extraction for main records, and the proportion of unresolved enrichment lookups. This enables users to evaluate data quality prior to conducting bibliometric, collaboration, geographic, or citation-network analyses.

BibFusion is distributed as a versioned Python package with explicit dependencies specified in requirements.txt, enabling environment recreation and consistent execution across systems. The pipeline is implemented through a designated entry point, such as "python run_main.py," adhering to a predetermined input/output directory structure that demarcates source-specific staging artifacts from the final integrated dataset. This directory convention includes directories designated for specific data sources, namely "Scopus_results/," "WoS_results/," and "all_data_wos_scopus/." The reproducibility of the process is further supported by deterministic file outputs (the seven entity tables plus audit artifacts) and log files that record enrichment and merge operations. In computational terms, runtime scales primarily with the number of records processed and with the optional OpenAlex enrichment step. DOI-based resolution is typically the main bottleneck and can be constrained by external API rate limits. As a result, large-scale runs may require batching, caching, or scheduled execution to ensure reliable completion. The version utilized in this study is BibFusion v1.0.0, which is available at the project repository: https://github.com/ladmepaz/bibfusion. The repository contains the source code, an example folder structure, and a versioned data dictionary (CSV/MD) that documents all output variables, including definitions, provenance (Scopus/WoS/derived), and normalization rules.

4. RESULTS

4.1. Integrated corpus produced by BibFusion

BibFusion produces a unified relational corpus organized into seven entities, as defined by the data model in Section 2.3: Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal, and Scimagodb. As illustrated in Table 2, the final sizes of each entity for the specified run, as outlined in Section 3.1, are reported. This table offers a concise summary of the integrated dataset that was generated by the pipeline.

Entity (CSV)	Rows	What it represents
Articles (All_Articles.csv)	8,569	Work records (main + reference records)
Authors (All_Authors.csv)	17,219	Disambiguated persons (unique PersonID)
ArticleAuthor (All_ArticleAuthor.csv)	28,254	Authorship links (work-author associations)
Citation (All_Citation.csv)	24,392	Directed citation edges (SR → SR_ref)
Affiliation (All_Affiliation.csv)	34,488	Work-author-affiliation instances (with country)
Journal (All_Journal.csv)	831	Normalized journals/sources (journal_id)
Scimagodb (All_Scimagodb.csv)	740	Journals with metrics (joinable by journal_id)

Table 2. Entity sizes of the unified corpus produced by BibFusion.
Note: The Articles entity includes 253 main records and 8,316 reference records parsed from cited references. Data generated by the authors.

In response to the query regarding entrepreneurial marketing, the Articles entity has been found to contain a total of 8,569 bibliographic records. This number is comprised of 253 main records that were retrieved directly from Scopus/WoS and 8,316 reference records that were parsed from cited-reference strings. The authorship layer consists of 17,219 disambiguated persons in Authors and 28,254 authorship links in ArticleAuthor, while the citation layer contains 24,392 directed citation edges in Citation. The remaining entities —Affiliation, Journal, and Scimagodb— capture the standardized institutional and geographic information, as well as the consolidated outlet metadata and metrics. These entities contextualize the unified corpus for downstream analyses.

4.2. Integration outcomes and dataset readiness (QC evidence)

As illustrated in Table 3, key integration and quality indicators are reported, demonstrating that the corpus is analysis-ready

following harmonization. At the level of main records —that is, publications retrieved directly from the query— BibFusion preserves cross-source provenance via sources_merged. This indicates that 183 out of 253 (72.3%) main records are present in both Scopus and WoS, while 59 out of 253 (23.3%) are Scopus-only and 11 out of 253 (4.3%) are WoS-only. This overlap indicates that, in the absence of integration, the combined raw retrieval would amount to 436 database-specific main records (Scopus: 242 and WoS: 194), accompanied by 183 cross-source duplicates. Through integration, BibFusion consolidates these into 253 unique main works, thereby preventing double-counting in downstream productivity, collaboration, and geographic indicators.

Indicator	Value
Unique main records (integrated)	253
Reference records (parsed from citations)	8,316
Main records by provenance (sources_merged)	183 (both); 59 (Scopus-only); 11 (WoS-only)
Raw main retrieval before merge (Scopus + WoS)	242 + 194 = 436
Cross-source duplicates consolidated (main)	183
DOI coverage (all Articles)	8,054/8,569 = *94%*
DOI coverage (main records)	220/253 = *87%*
DOI coverage (matched main in both sources)	181/183 = *98.9%*
Country completeness in Articles (main records)	252/253 = *99.6%*
Affiliation rows with country populated	34,488/34,488 = *100%*
Citation edges (unique)	24,392 (0 duplicate edges)
Citation edges originating from main records	23,927/24,392 = *98.1%*
Citing endpoint resolvable in Articles (SR)	24,392/24,392 = *100%*
Cited endpoint resolvable in Articles (SR_ref)	13,275/24,392 = *54.4%*
Journals linked to Scimagodb metrics (journal_id)	727/831 = *87.5%*
Temporal coverage	Main: 2005-2025; References: 1900-2026

Table 3. Integration outcomes and quality indicators supporting dataset readiness.
Note: The cited endpoint resolvability rate is expected to be <100% because many cited items (e.g., books or out-of-coverage sources) do not appear as fully described records in All_Articles; edges are retained using the normalized SR_ref key. Data generated by the software Python.

The completeness of identifiers and metadata further supports the dataset's readiness. The overall DOI coverage was found to be 94% (8,054 out of 8,569), with a 87% coverage for main records (220 out of 253). A strong complementarity effect was observed across sources. The proportion of DOI coverage among matched main records is 98.9% (181 out of 183), while it is lower for source-unique main records (Scopus-only: 57.6% and WoS-only: 45.5%), indicating that merging substantially improves identifier availability for the unified corpus. The geographic metadata for the main corpus is found to be highly complete. The article-level country field is populated for 252 out of 253 (99.6%) main records, and the Affiliation entity provides a normalized country value for 100% of affiliation rows. Affiliations are available for 251 out of 253 (99.2%) main records. Outlet consolidation is also robust. A total of 87.5% of normalized journals (727 out of 831) are linked to Scimagodb metrics via journal_id, thereby enabling stratified analyses by journal indicators. In conclusion, the use of citation linking has been demonstrated to exhibit a reliable structure for network construction while maintaining transparency regarding coverage limits. The citation table contains 24,392 directed edges, with no duplicate edges under the (SR, SR_ref) key. All citing keys (SR) resolve to the Articles entity (100%), and 98.1% of citation edges originate from main records. For the specified endpoint, 54.4% of SR_ref values are mapped to an Articles record, while the remaining values correspond to out-of-coverage items, such as books or sources not represented as full records. This process generates a citation network that is

both functional and explicit. That is to say, the edges are preserved and standardized, and the cited nodes are reference-only keys rather than fully described works.

5. DISCUSSION

5.1. Implications

The integration and QC evidence reported in Tables 2 and 3 directly impacts the validity of downstream scientometric analyses. First, the consolidation of cross-source duplicates and the preservation of provenance (sources_merged) serve to prevent double-counting and stabilize productivity and trend indicators. In the absence of deduplication, the same work can appear as two separate records, resulting in inflated publication counts and distorted time series patterns and outlet rankings. Second, the enhancement of identifier coverage, particularly the near-complete availability of DOIs across records aligned between Scopus and WoS, fortifies record linkage and facilitates more dependable connections to external resources, thereby mitigating fragmentation in citation and metadata enrichment workflows. Third, the high completeness of country extraction for main records (and full country population in the Affiliation table) materially improves scientific geography analyses by reducing "unknown country" noise and allowing country-level production and collaboration maps to reflect institutional signals rather than missing data artifacts. In conclusion, the utilization of the canonical PersonID in conjunction with the explicit authorship link table facilitates the establishment of more credible coauthorship and collaboration networks. This enhancement is achieved by mitigating the occurrence of author splitting or merging that is precipitated by orthographic variants, including but not limited to accents, initials, and the sequence of authorship. When considered as a whole, these QC enhancements progress the dataset from a collection of heterogeneous exports to a traceable relational corpus. In this new form, bibliometric indicators, country-based comparisons, and network measures can be computed reproducibly with a reduced risk of bias driven by metadata inconsistencies (Crystal-Ornelas *et al.,* 2022).

5.2. Positioning and limitations

Conceptually, BibFusion occupies a middle ground between general-purpose bibliometric toolkits and ad hoc, project-specific integration scripts. Many existing workflows facilitate the importation of Scopus and WoS exports into a unified flat table for descriptive indicators. However, these workflows frequently result in the transformation of cited references into database-specific strings, thereby failing to establish a shared, normalized reference-key space across sources. Consequently, the construction of citation networks and the deduplication of references can become fragile and challenging to reproduce. BibFusion addresses this gap by materializing the integrated corpus as an explicit ER model and by generating dedicated linking tables (ArticleAuthor, Citation, Affiliation) anchored to canonical identifiers (SR and PersonID). Provenance is preserved through the use of sources_merged, and integration decisions are externalized via audit artifacts to support inspection and reruns. In practice, this design reduces cross-source double counting, stabilizes author- and country-level attribution for the main corpus, and produces an analysis-ready citation edge list with clear, transparent boundaries between resolvable and out-of-coverage cited items.

Concurrently, BibFusion does not eliminate the intrinsic constraints of bibliographic metadata (Visser *et al.,* 2021). Cross-source integration is predicated on persistent identifiers, and thus a DOI-first strategy is vulnerable to missing, malformed, or inconsistently recorded DOIs. Although BibFusion applies conservative fallback rules when a DOI is unavailable, any non-DOI matching retains a residual risk of both missed matches and occasional false merges. The use of ORCID and OpenAlex identifiers in conjunction with name normalization serves to strengthen author disambiguation. However, cases of homonymy and incomplete identifier coverage can still result in ambiguous situations that necessitate human review. Therefore, the generation of explicit conflict files and alias logs is preferable to the imposition of deterministic decisions. Similarly, citation linking is inherently incomplete when cited items fall outside the scope of the exported databases (e.g., books, reports, or non-indexed

sources). BibFusion preserves these relations as reference-keyed nodes but cannot guarantee complete metadata for out-of-coverage works. In this regard, the contribution of BibFusion does not lie in the resolution of imperfect metadata, but rather in the provision of a reproducible, audit-ready integration workflow. This workflow serves to mitigate avoidable inconsistencies and renders residual uncertainty evident for rigorous downstream scientometric analysis.

5.3. Practical guidance

BibFusion's relational outputs are structured to support common scientometric analyses in a straightforward and reproducible way. For the purpose of trend and productivity analyses, the Articles table functions as the primary unit of analysis. Typically, the filter is set to ismainarticle = True, allowing for the computation of publication time series (by year) and outlet-level summaries (via source_title/journal) on the deduplicated corpus without the inflation of counts. Additionally, the sources_merged field facilitates sensitivity checks by source provenance. In the context of scientific geography, country-level results can be obtained at two complementary resolutions. The first is a lightweight view, which utilizes the article-level country field in Articles for the purpose of main-corpus production maps. The second is a finer-grained view, which employs the Affiliation table to analyze author-institution-country linkages. Examples of such linkages include multicountry outputs, institutional collaboration patterns, and fractional counting approaches (Demaine, 2022; Hottenrott *et al.,* 2021; Matveeva *et al.,* 2022; Sivertsen *et al.,* 2025). For the purpose of coauthorship analyses, ArticleAuthor employs a normalized bipartite structure to link works (SR) to disambiguated authors (PersonID). This representation enables the derivation of coauthor networks through the projection of author-author ties within each SR, with the option of assigning edge weights based on the number of shared publications. For citation-based analyses, Citation provides a cleaned directed edge list (SR → SR_ref) that supports citation graphs, basic impact measures (in-degree/out-degree), and co-citation analyses (Chen *et al.,* 2024; Wang *et al.,* 2023; Yang *et al.,* 2024). When a cited key (SR_ref) resolves to a record in Articles, cited nodes can be enriched with full metadata; when it does not, the edge is retained with the normalized reference key, so the network structure remains intact while clearly marking out-of-coverage items. In conclusion, Journal and Scimagodb contextualize results at the outlet level by linking publications to normalized journal identities and, when available, to indicators such as SJR quartiles and subject categories, enabling consistent stratified reporting across the unified corpus (Lim *et al.,* 2024; Schmal, 2024; Vaccaro *et al.,* 2022).

To ensure the attainment of consistent results, it is imperative that BibFusion be executed on complete exports from both databases, employing standardized settings. For Scopus, records should be exported as a CSV file with the "Select all information" option enabled. For WoS, records should be exported as a TXT file with the "Full Record and Cited References" option selected, as outlined in Appendix B. In both cases, users should avoid truncation options intended for spreadsheet compatibility (e.g., "optimize for Excel"), as these can shorten long fields such as affiliations, abstracts, and reference strings and degrade matching and citation linking. In addition, it is recommended to maintain a consistent search strategy across various sources, ensuring alignment in terms of fields, years, languages, and document types. To ensure comprehensive reproducibility, it is essential to archive precise queries and export dates alongside the raw files. During execution, it is considered best practice to preserve the default folder structure (i.e., Scopus_results/, WoS_results/, and all_data_wos_scopus/) and to retain all generated logs and audit artifacts. It is imperative that users thoroughly examine audit outputs, with a particular focus on alias/conflict files and merge logs, prior to conducting analyses at the author- or institution-level. Furthermore, substantive indicators such as trends, geography, and collaboration should be restricted to main records (ismainarticle = True), unless the objective of the study is to examine the reference layer explicitly. For large corpora, optional OpenAlex enrichment should be planned meticulously: It is important to note that DOI resolution can be subject to rate limitations. Consequently, when attempting to

scale up, the implementation of batching and caching is strongly advised. In conclusion, downstream analyses should rely on the relational joins defined by the data model (SR, PersonID, journal_id) rather than re-parsing raw strings. This will ensure that results remain consistent, auditable, and reproducible across runs.

6. CONCLUSION

This study presented BibFusion, a reproducible preprocessing pipeline that harmonizes Scopus and WoS exports into a unified, traceable, analysis-ready corpus. BibFusion operationalizes a unified ER model and produces seven relational tables —Articles, Authors, ArticleAuthor, Citation, Affiliation, Journal, and Scimagodb— supported by explicit primary/foreign keys and provenance fields to preserve relational integrity across entities. The pipeline implements systematic normalization, which includes ASCII/uppercase standardization and SR/SR_ref cleaning. It also implements conservative cross-source integration, which includes DOI-first deduplication with controlled fallbacks. The pipeline can also identify the canonical author via PersonID and consolidate citations and outlets in a structured manner. A fundamental objective is the integration of relational structure with auditability: BibFusion produces standardized outputs for bibliometric and network analyses, as well as audit artifacts (i.e., aliases, potential conflicts, and merge logs) that facilitate transparency in key decisions and enable targeted human review when metadata remains ambiguous. The integration of these components establishes a pragmatic foundation for reproducible scientometric workflows, thereby mitigating avoidable inconsistencies across sources and explicitly documenting residual uncertainty rather than concealing it. BibFusion provides practical value to the scientometrics and bibliometrics community by transforming two heterogeneous, widely used data sources into a reproducible and auditable integration workflow. By standardizing identifiers, consolidating duplicates, and exposing a clear ER structure, the package reduces the repetitive, error-prone manual curation that often precedes bibliometric analyses —particularly for collaboration and scientific geography studies where author and affiliation inconsistencies can strongly bias results. Furthermore, BibFusion's incorporation of audit logs and conflict/alias artifacts serves to make residual uncertainty explicit, thereby enabling transparent inspection and correction without compromising the reproducibility of results. Consequently, researchers can allocate less effort toward ad hoc cleaning and more toward interpretable analyses, while maintaining a documented trail that supports verification, extension, and reuse across studies and research domains.

It has been demonstrated that the implementation of multiple extensions has the potential to enhance the functionality of BibFusion and expand its range of applications. First, the robustness of matching could be enhanced by incorporating supplementary similarity signals that extend beyond the current conservative cascade. These additional signals might include structured field agreement on journal/volume/pages, calibrated string-similarity thresholds, and probabilistic or learned record-linkage models. This approach would ensure the preservation of auditability and the control of false merges. Second, the scope of coverage could be expanded by incorporating supplementary sources and formats, such as OpenAlex snapshots, Crossref metadata, dimensions, PubMed, or institutional repositories. Additionally, enhancing the ingestion process to align with evolving export schemas could further expand the scope of coverage. Third, the efficacy of quality assurance could be augmented through the implementation of automated QC dashboards that facilitate the reporting of readiness metrics (e.g., DOI and country completeness, duplicate consolidation rates, citation endpoint resolvability, and author ID coverage) and the identification of high-impact ambiguity clusters for subsequent review. In conclusion, workflow scalability can be enhanced through caching and incremental updates (e.g., re-running only enrichment or only merging on newly added records), facilitating efficient maintenance of longitudinal corpora as databases and identifiers evolve over time.

Acknowledgments

The authors acknowledge the support of the Universidad Nacional de Colombia, MetricSci,

and the Data Lab at the Universidad Nacional de Colombia, Sede de La Paz, for facilitating the development and testing of the BibFusion preprocessing pipeline, providing repository hosting and computational equipment, and supporting students involved in this project. The authors also acknowledge OpenAI for providing AI tools used during the development and writing process (e.g., Codex and ChatGPT) to assist with generating and optimizing portions of the codebase and to support manuscript organization and iterative revision. All outputs produced with AI assistance were reviewed and validated by the authors, who accept full responsibility for the final code and manuscript.

Funding

This research received funding from Universidad Nacional de Colombia, Sede de La Paz, through the project "Estrategias Innovadoras de Marketing Emprendedor: Un Caso Aplicado al Laboratorio de Datos de la Universidad Nacional" (Hermes 64027).

Conflict of interest

The authors declare no conflicts of interest.

Contribution statement

Angelo Britto: Conceptualization, software, data curation, methodology, validation, writing – review & editing.

Sebastian Robledo: Conceptualization, software, methodology, data curation, formal analysis, visualization, writing – original draft, writing – review & editing, project administration.

Martha Zuluaga: Validation, investigation, formal analysis, visualization, writing – review & editing.

Statement of data consent

The bibliographic data utilized in this study were obtained through licensed exports from Scopus and Web of Science and are subject to the terms and conditions of the respective providers. Consequently, the exported files in their raw state are not disseminated publicly. The processed outputs generated by BibFusion (aggregated CSV entity tables) and the code/workflow details can be made available upon reasonable request, subject to compliance with data provider licensing restrictions.

REFERENCES

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

APPENDICES

Appendix A. Source-specific workflow diagrams

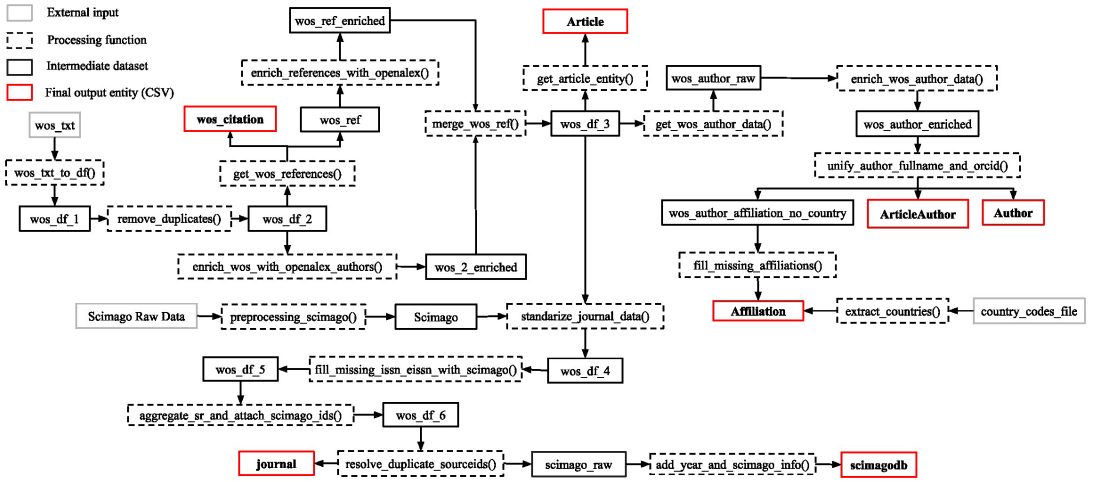

Figure A1. Function-level workflow for WoS TXT preprocessing. Note: This diagram provides a function-level view of the transformations applied to the WoS TXT export, complementing the end-to-end workflow in Figure 1. It traces the sequence from ingestion and canonicalization through metadata normalization (e.g., title and SR key standardization), affiliation parsing with country extraction, and the generation of standardized staging outputs used in the subsequent cross-database merge. Red-outlined boxes denote the standardized entity outputs that correspond to the unified ER model in Figure 2.

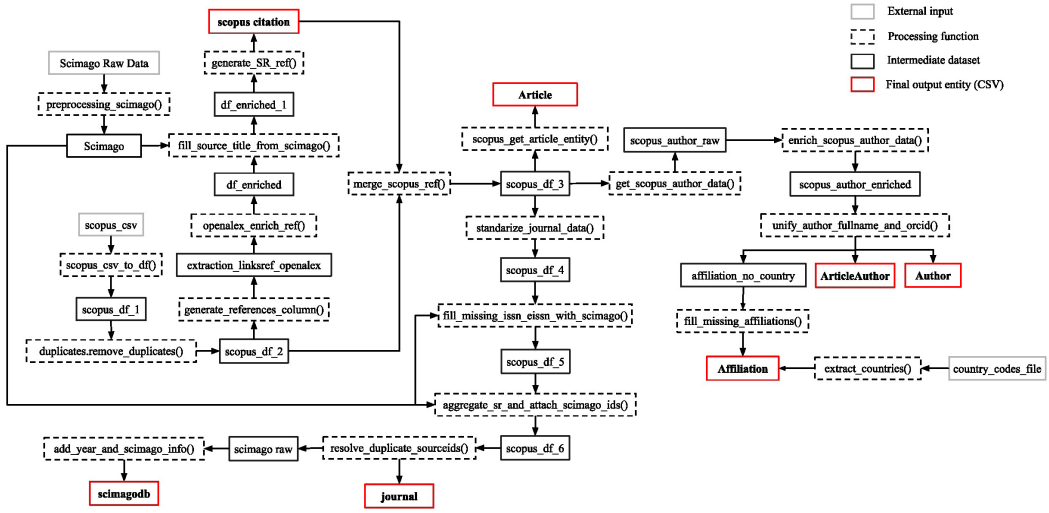

Figure A2. Function-level workflow for Scopus CSV preprocessing and DOI-based OpenAlex enrichment. Note: This diagram complements Figure 1 by detailing the sequence of functions applied to Scopus CSV exports. It shows ingestion and field normalization (including DOI standardization), DOI-driven enrichment via OpenAlex (e.g., work identifiers and author identifiers when available), and downstream processing steps such as reference/SR construction, journal/Scimago consolidation, and affiliation parsing with country extraction. The workflow produces standardized, source-aligned staging entities that feed the subsequent cross-database deduplication and merge step.

Appendix B. Export settings used for data collection (Scopus and WoS)

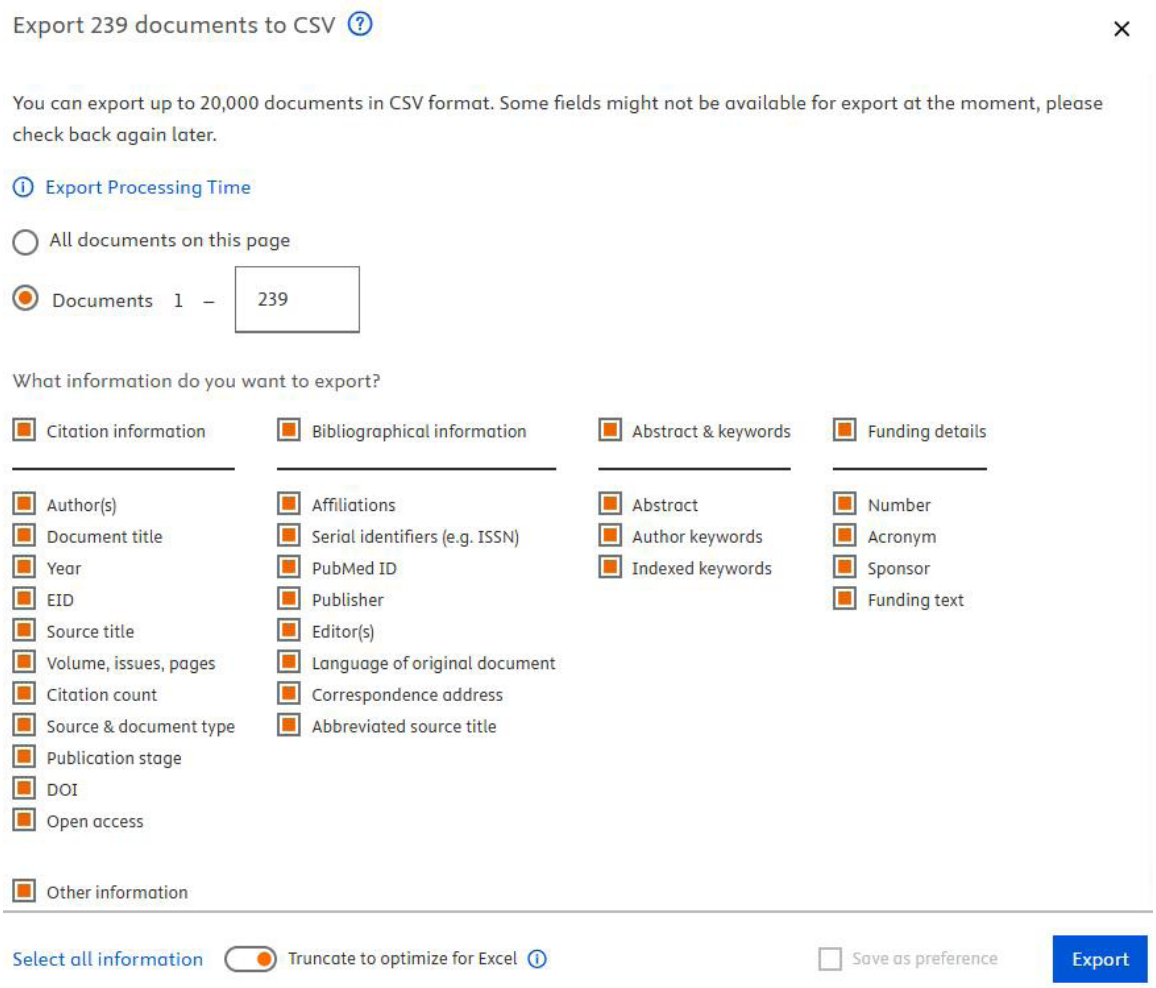

Figure B1. Scopus CSV export configuration ("Select all information"). Note: This figure documents the Scopus export settings used to generate the raw CSV input for BibFusion, in which the export is configured to include all available information (e.g., citation information, bibliographical information, abstracts/keywords, funding details, and other descriptors). These settings are required to preserve complete metadata and ensure consistent downstream normalization, deduplication, and traceability.

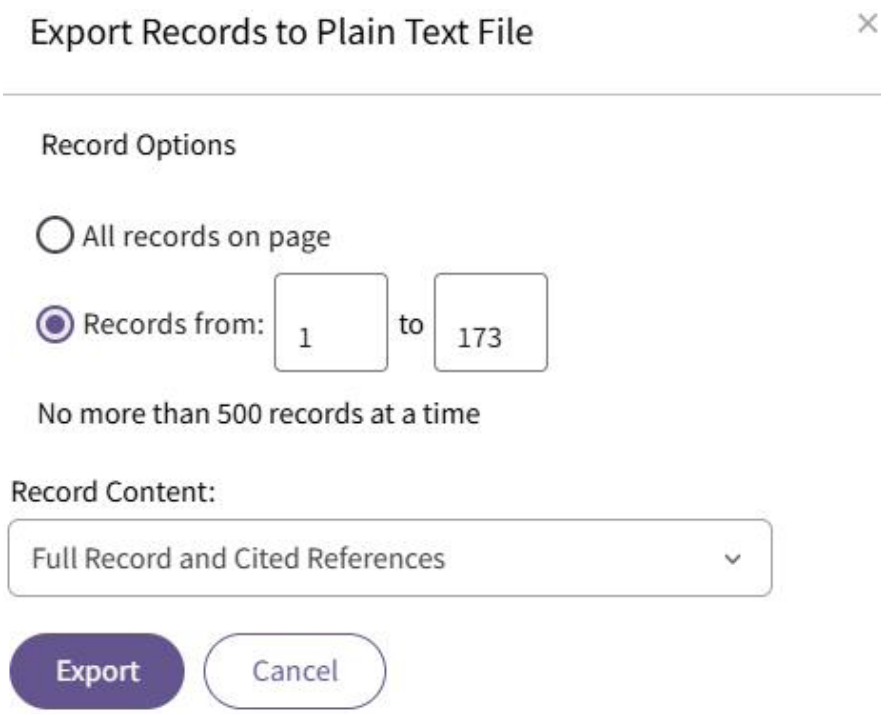

Figure B2. WoS TXT export configuration ("Full Record and Cited References"). Note: This figure documents the WoS export settings used to generate the raw TXT input for BibFusion, in which records are exported as a plain-text file with "Full Record and Cited References" selected. This configuration ensures that both full bibliographic metadata and cited-reference strings are available for SR/SR_ref normalization and citation-link construction in the unified corpus.