ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1957

redalycR: Automating bibliometric analysis in R through a reproducible architecture

Henry Osorto[1,2]
[1]Universidad Nacional Autónoma de Honduras, UNAH, Honduras, *henry.osorto@unah.edu.hn*
[2]Facultad de Postgrado, Universidad Tecnológica Centroamericana, UNITEC, Honduras, *henry.osorto@unitec.edu.hn*

Abstract– ***Th Literature review is a core component of the scientific research process; however, the rapid growth of academic production has exceeded the capacity of traditional manual approaches. In response to this challenge, computational tools oriented toward automation and reproducibility have gained increasing relevance in bibliometric analysis. In this context, this paper introduces redalycR, an R package designed to automate bibliometric analysis through a reproducible architecture based exclusively on open data from the Redalyc database. The paper describes the conceptual design and implementation of the package, as well as its integration into analytical workflows in R for the acquisition, cleaning, normalization, and structuring of bibliographic metadata. Through an illustrative application case, it is shown how redalycR enables the programmatic and reproducible generation of descriptive keyword analyses and keyword co-occurrence networks. The results demonstrate that the package facilitates the structural exploration of Spanish-language scientific corpora, complementing existing bibliometric tools and expanding analytical capabilities over open-access literature. Finally, the current limitations of the package are discussed, and a future development agenda aimed at its extension and strengthening is outlined.***

Keywords-- ***Bibliometrics; Text mining; Open science; R; Redalyc.***

redalycR: automatización del análisis bibliométrico en R mediante una arquitectura reproducible

Henry Osorto[1,2]
[1]Universidad Nacional Autónoma de Honduras, UNAH, Honduras, *henry.osorto@unah.edu.hn*
[2]Facultad de Postgrado, Universidad Tecnológica Centroamericana, UNITEC, Honduras, *henry.osorto@unitec.edu.hn*

Resumen– ***La revisión de literatura es un componente central del proceso de investigación científica; sin embargo, el crecimiento acelerado de la producción académica ha desbordado la capacidad de los enfoques manuales tradicionales. En respuesta a este desafío, las herramientas computacionales orientadas a la automatización y reproducibilidad han adquirido un papel cada vez más relevante en el análisis bibliométrico. En este contexto, el presente trabajo introduce redalycR, un paquete en R diseñado para automatizar el análisis bibliométrico mediante una arquitectura reproducible basada exclusivamente en datos abiertos provenientes de la base de datos Redalyc. El artículo describe el diseño conceptual y la implementación del paquete, así como su integración en flujos analíticos en R para la adquisición, limpieza, normalización y estructuración de metadatos bibliográficos. A través de un caso de aplicación ilustrativo, se muestra cómo redalycR permite generar análisis descriptivos de palabras clave y redes de co-ocurrencia de forma programática y reproducible. Los resultados evidencian que el paquete facilita la exploración estructural de corpus científicos en español, complementando herramientas bibliométricas existentes y ampliando las posibilidades de análisis sobre literatura de acceso abierto. Finalmente, se discuten las limitaciones actuales del paquete y se plantea una agenda de desarrollo futuro orientada a su extensión y fortalecimiento.***

Palabras clave-- Bibliometría; Minería de texto; Ciencia abierta; R; Redalyc.

I. Introducción

La revisión de literatura constituye un componente central del proceso de investigación científica, en tanto permite construir el marco conceptual y teórico del estudio y sustentar decisiones metodológicas. En el ámbito formativo y de investigación, se ha señalado que una revisión de literatura rigurosa es condición necesaria para elevar la sofisticación metodológica y la utilidad del conocimiento producido, especialmente en contextos doctorales [1]. Sin embargo, el crecimiento acelerado de las publicaciones científicas y su digitalización han ampliado la escala del problema: hoy se gestiona un universo de información con millones de autores, artículos, citas y redes académicas que demandan enfoques computacionales para su manejo y análisis [2].

En este escenario, la automatización y el uso de minería de texto han adquirido relevancia como respuesta a los límites prácticos de la revisión manual. En particular, se han propuesto marcos y sistemas para construir y evaluar revisiones automatizadas, destacando la necesidad de contar con metodologías reproducibles y mecanismos objetivos de evaluación, por ejemplo mediante el uso de listas de referencias de revisiones previas como "ground truth" para entrenar y comparar modelos [3]. De forma complementaria, revisiones de aplicaciones de minería de datos en bibliotecas académicas han mostrado cómo técnicas como agrupamiento, clasificación, asociación y regresión se aplican a servicios, calidad, colecciones y comportamiento de uso, evidenciando el valor de enfoques computacionales para organizar y filtrar grandes volúmenes de información [4]. Esta literatura, en conjunto, converge en un punto: a mayor escala de producción científica, mayor necesidad de procesos automatizables, verificables y reutilizables.

El desarrollo de infraestructura de software ha sido determinante para operacionalizar estas estrategias. En el ecosistema de R, el paquete tm aportó tempranamente una infraestructura para tareas típicas de minería de texto análisis basado en conteos, clustering, clasificación y kernels y consolidó un marco reusable para investigaciones que requieren procesamiento textual [5]. Paralelamente, la evolución de enfoques bibliométricos y su formalización como campo incluyendo análisis de citación y sus fundamentos conceptuales ha sido discutida desde perspectivas que sistematizan teorías, técnicas y aplicaciones, reforzando la pertinencia de enfoques cuantitativos para estudiar la ciencia [6].

Desde el punto de vista metodológico, múltiples trabajos han evidenciado que el análisis automatizado permite caracterizar comunidades científicas a gran escala mediante minería de textos y modelado de tópicos. En ingeniería de software, por ejemplo, se ha mostrado que los métodos automáticos hacen "rápidamente repetible" el hallazgo de tendencias temáticas en decenas de miles de artículos y facilitan que otros investigadores repliquen el análisis bajo nuevos supuestos o nuevos datos [7]. En una línea similar, estudios bibliométricos de gran alcance han utilizado análisis automatizado de citaciones y modelado de tópicos para caracterizar campos con decenas de miles de artículos, resaltando explícitamente la conveniencia de enfoques automatizados cuando no es viable realizar revisiones sistemáticas completas por restricciones de tiempo o por ausencia de preguntas de investigación altamente específicas [8]. Adicionalmente, análisis de subconjuntos de literatura han mostrado que el enfoque es transferible a otros corpus y escalable para estudios masivos [9]. Fuera del dominio de ingeniería de software, también se han propuesto marcos genéricos de minería de texto automatizada para examinar tendencias en horizontes temporales amplios, validando

resultados contra revisiones previas y combinando bibliometría con meta-revisión [10].

Más recientemente, la automatización se ha expandido hacia enfoques híbridos que combinan bibliometría con técnicas modernas de inteligencia artificial. Un ejemplo es EmbedSLR, un framework open-source que estandariza el "screening" de revisiones sistemáticas mediante embeddings y distancia coseno, complementándolo con auditoría bibliométrica, y enfatizando su carácter determinista y reproducible como ventaja frente a procesos lentos y propensos a sesgos [11]. En paralelo, la aparición de grandes modelos de lenguaje (LLMs) ha impulsado investigaciones que evalúan su desempeño en tareas de extracción, agregación y razonamiento contextual con datasets curados y marcos de evaluación reproducibles [12]. Esta convergencia refuerza el argumento de que los flujos reproducibles ya no solo deben automatizar la descarga y limpieza de datos, sino también facilitar integraciones futuras con componentes de IA, manteniendo control metodológico y trazabilidad.

En el ámbito aplicado, los estudios bibliométricos continúan expandiéndose hacia dominios diversos y muestran que el análisis automatizado se consolida como práctica transversal: se observan revisiones sistemáticas acompañadas de bibliometría para mapear campos emergentes [13]; análisis bibliométricos para comprender paisajes globales de investigación, como IA en educación usando herramientas de mapeo científico [14]; y estudios bibliométricos basados en R para explorar la integración de IA en economía colaborativa y detectar temas emergentes [15]. Asimismo, revisiones apoyadas en VOSviewer han aplicado minería de texto y clustering para comparar evoluciones temáticas [16]. Aunque algunos de estos trabajos son dominio-específicos, su valor para este artículo radica en mostrar cómo las comunidades científicas están adoptando análisis bibliométrico+minería de texto como mecanismo estándar para organizar evidencia en contextos de alta producción.

Un elemento contextual clave que habilita este tipo de enfoques es la creciente apertura de infraestructuras de datos científicos bajo principios de ciencia abierta. En este sentido, Redalyc ha desarrollado y puesto a disposición mecanismos formales para la reutilización de metadatos y contenidos académicos, incluyendo su rol como data provider bajo el protocolo OAI-PMH, el desarrollo de un API para la descarga y consulta de metadatos reutilizables, y la disponibilidad de contenidos en formato XML JATS con fines de minería de datos. Estas iniciativas, orientadas a aumentar la visibilidad y reutilización de revistas de acceso abierto diamante, reducen significativamente las barreras técnicas para el desarrollo de flujos computacionales reproducibles basados en fuentes no propietarias.

En esta línea de investigación, el antecedente directo de este trabajo es el artículo presentado en LACCEI 2025, donde se propuso y aplicó un enfoque automatizado de revisión de literatura mediante técnicas de text mining en R, utilizando metadatos masivos extraídos desde una fuente abierta y destacando la replicabilidad del procedimiento mediante código [17]. No obstante, una limitación natural de los enfoques presentados como "sintaxis" o scripts es que suelen permanecer como implementaciones puntuales, con fricción para su reutilización, documentación, control de dependencias y extensibilidad. En otras palabras, aunque los scripts pueden demostrar factibilidad, la consolidación en forma de paquete es un paso metodológico y de ingeniería de software que incrementa la reutilización, el mantenimiento, la estandarización y la integración con flujos bibliométricos más amplios.

Por ello, este artículo presenta redalycR, un paquete en R orientado a automatizar el análisis bibliométrico mediante una arquitectura reproducible y modular, facilitando la adquisición, limpieza, normalización y estructuración de metadatos de la base de datos de Redalyc para su integración con flujos analíticos en R. El enfoque enfatiza prácticas de software científico abierto, alineadas con experiencias previas de paquetización para facilitar el acceso y la integración de grandes colecciones bibliográficas en pipelines reproducibles [18]. Además, el diseño del paquete busca ser extensible para incorporar nuevas variables y capacidades analíticas conforme se disponga de mayor granularidad de datos y/o acuerdos institucionales que habiliten enriquecimiento de metadatos, manteniendo como objetivo transversal la reproducibilidad y escalabilidad del proceso.

En ese sentido, las principales contribuciones de este trabajo son:

- El diseño de una arquitectura reproducible para automatizar la recuperación y procesamiento de metadatos desde Redalyc;
- La implementación de un paquete en R funcional y disponible en GitHub;
- La adaptación de los metadatos de Redalyc a una estructura bibliométrica compatible con flujos en R;
- La generación automatizada de análisis de frecuencia y redes de co-ocurrencia de palabras clave;
- La demostración de un flujo reproducible aplicado a literatura científica en español.

II. Materiales y Métodos

A. *Fuente de datos*

El pipeline implementado por redalycR opera exclusivamente con registros recuperados desde Redalyc, de modo que el corpus de análisis se define por una búsqueda particular ejecutada sobre esta fuente. El objetivo metodológico es facilitar: (i) la obtención reproducible del corpus; (ii) su transformación a una estructura bibliométrica utilizable por flujos típicos en R; y (iii) la generación de análisis exploratorios de palabras clave y redes de co-ocurrencia.

B. *Software*

El desarrollo del paquete redalycR fue realizado íntegramente en el lenguaje estadístico R [19], aprovechando su ecosistema orientado al análisis reproducible y al desarrollo de

software científico abierto. El código fuente del paquete se encuentra disponible públicamente en GitHub, en el repositorio "Henry-Osorto/redalycR", lo que permite la inspección, reutilización y extensión del software por parte de la comunidad académica, así como el control de versiones y la trazabilidad del proceso de desarrollo.

Para desarrollar las funciones se utilizaron las siguientes librerías: httr2 [20], jsonlite [21], tibble [22], dplyr [23], tidyr [24], purrr [25], stringr [26], ggplot2 [27], digest [28], cli [29], igraph [30], Matrix [31], cld2 [32], cld3 [33], writexl [34], y svglite [35].

C. *Versión del paquete*

Para garantizar la reproducibilidad del caso de estudio, se documentaron la versión de R, la versión del paquete, el repositorio de GitHub, y la fecha en la que comenzó a estar disponible de acceso público. Esta información permite replicar el ambiente mínimo de ejecución y controlar posibles variaciones derivadas de cambios posteriores en el paquete o en la fuente de datos.

TABLA I
DESCRIPCIÓN DEL PAQUETE REDALYCR

Elemento	Descripción
Lenguaje de programación	R version 4.5.3
Nombre del paquete	redalycR
Versión del paquete	0.1.1
Repositorio	GitHub: Henry-Osorto/redalycR
Sistema operativo	Windows 10/11, macOS o Linux
Disponible desde	27 de enero de 2026
Licencia	MIT + file LICENSE

D. *Arquitectura de los módulos*

La arquitectura de redalycR se organiza en módulos funcionales que corresponden a etapas típicas de un flujo bibliométrico reproducible. En la Figura , se visualiza el diagrama de flujo de la arquitectura del paquete redalycR.

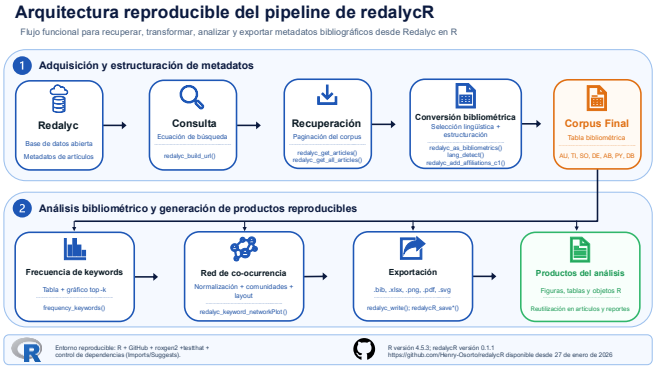

Fig. 1 Diagrama de flujo de la arquitectura del paquete redalycR

1) Definición de consulta, construcción de URL y obtención de los datos: este módulo estandariza la manera en que se construye la consulta hacia la API de Redalyc. Las funciones de este módulo son:

- redalyc_build_url: construye de forma parametrizada la URL/consulta para extracción.
- redalyc_get_articles: obtiene artículos a partir de una consulta definida.
- redalyc_get_all_articles: automatiza la recuperación masiva (paginación o extracción exhaustiva según el criterio implementado).

2) Normalización, enriquecimiento y estandarización bibliométrica: Este módulo transforma los resultados crudos en una estructura coherente para análisis bibliométrico y mapeo científico. Se prioriza la compatibilidad con convenciones bibliográficas comunes, de modo que la salida pueda integrarse en flujos existentes. Funciones:

- redalyc_as_bibliometrics: convierte el resultado a una estructura bibliométrica estandarizada.
- redalyc_add_affiliations_c1: construye variables de afiliación y la columna C1 (autor–afiliación) para aproximarse a estructuras tipo Scopus/WoS cuando la fuente lo permite.
- redalyc_lang_detect: apoya decisiones de normalización lingüística (selección/etiquetado de idioma cuando aplica).

3) Preparación de texto y normalización de palabras clave: Este módulo se orienta a asegurar consistencia en variables textuales (resúmenes, títulos, keywords), que suelen contener ruido (variantes ortográficas, separadores heterogéneos, duplicados, mezcla de idiomas). Funciones:

- clean_keyword: normaliza entradas de palabras clave (limpieza, estandarización de formato).
- split_keywords: separa campos multivaluados (keywords) usando reglas de separación definidas.

4) *Exportación, persistencia e interoperabilidad:* Este módulo facilita la integración con herramientas bibliométricas y el resguardo sistemático de productos de investigación (tablas, figuras, archivos de intercambio). Funciones:

- redalyc_write_bib: exporta resultados a un formato bibliográfico.
- redalycR_save_xlsx: guarda salidas tabulares en Excel.
- redalycR_save_plot: guarda gráficas generadas por funciones del paquete.
- redalycR_output_dir: estandariza la lógica de directorios de salida para garantizar la organización de los outputs.

5) Mapeo científico: este módulo implementa funciones orientadas a generar productos típicos del análisis bibliométrico (frecuencia y redes de co-ocurrencia de keywords), priorizando parámetros reproducibles y guardado opcional. Funciones:

- frequency_keywords: calcula y visualiza frecuencias de términos/keywords como etapa de análisis exploratorio.
- keyword_networkPlot: construye/visualiza redes de co-ocurrencia de keywords a partir de datos preparados.

E. Método de red de co-ocurrencia y normalización

Para construir la red de co-ocurrencia se extrae la columna de keywords (DE) y se transforma a vector de texto, luego se separan las palabras y se normalizan removiendo duplicados por documento y descartando vacíos. Luego se calcula la frecuencia documental de cada término como conteo de documentos donde aparece. Se aplica un filtro de términos (min_freq = 2 por defecto).

Con el conjunto filtrado de términos, se construye una matriz dispersa documento-término $\boldsymbol{X}$ (valores binarios 0/1) y se calcula la matriz de co-ocurrencia:

$$\mathbf{C} = \mathbf{X}^{\top}\mathbf{X}, \text{ y se fuerza diag}(\mathbf{C}) = \mathbf{0}$$

donde c_{ij} representa el número de documentos en los que co-ocurren los términos i y j. Para un par de términos i y j.

Se define además c_i como la frecuencia del término i que corresponde a la diagonal de $\mathbf{X}^{\top}\mathbf{X}$ antes de anularla.

Si el parámetro normalize no es NULL, la función aplica una normalización sobre $\mathbf{C}$ para producir $\mathbf{W}$ (matriz ponderada), usando las siguientes definiciones:

Association: $$w_{ij} = \frac{c_{ij}}{c_i c_j} \quad (1)$$

Jaccard: $$w_{ij} = \frac{c_{ij}}{c_i + c_j - c_{ij}} \quad (2)$$

Inclusion: $$w_{ij} = \frac{c_{ij}}{\min(c_i, c_j)} \quad (3)$$

Salton: $$w_{ij} = \frac{c_{ij}}{\sqrt{c_i c_j}} \quad (4)$$

Equivalence: $$w_{ij} = \frac{c_{ij}^2}{c_i c_j} \quad (5)$$

III. Resultados

Como parte de los resultados metodológicos de este estudio, se presenta la implementación práctica del paquete redalycR, el cual se encuentra disponible públicamente en un repositorio de GitHub. La instalación del paquete se realiza directamente desde dicho repositorio, utilizando herramientas estándar del ecosistema R para la gestión de software científico reproducible. La instalación puede efectuarse mediante el siguiente comando:

```
# Instalación desde GitHub
# install.packages("remotes")
remotes::install_github("Henry-Osorto/redalycR")
```

Una vez instalado, el paquete se carga en la sesión de trabajo de R de forma habitual: *library(redalycR)*. Este mecanismo de distribución permite que los usuarios accedan a la versión más reciente del código, facilita la incorporación de mejoras continuas y respalda la transparencia del desarrollo del software, en línea con prácticas de ciencia abierta y software reproducible.

Para ilustrar el funcionamiento del paquete, se ejecutó un ejemplo aplicado basado en una consulta temática específica relacionada con *capital humano* y *desarrollo económico*. La consulta se definió explícitamente como:

```
library(redalycR)
query <- "\"capital humano\" AND \"desarrollo economico\""
```

A partir de esta consulta, se recuperó el conjunto completo de artículos disponibles en Redalyc que cumplen con el criterio de búsqueda, utilizando la función redalyc_get_all_articles(). Posteriormente, los metadatos obtenidos fueron transformados a una estructura bibliométrica homogénea mediante la función redalyc_as_bibliometrics(), priorizando el idioma español para los campos textuales:

```
df <- redalyc_get_all_articles(query, page_size = 100)
df.redalyc <- redalyc_as_bibliometrics(df, lang = "es")
```

La Tabla 2 describe el resultado de la consulta de la obtención de un corpus de un caso demostrativo que ejemplifica el uso del paquete.

TABLA II
CARACTERIZACIÓN DEL CORPUS UTILIZADO EN EL CASO DE ESTUDIO DE REDALYCR

Elemento	Descripción
Fuente	Redalyc
Fecha de consulta	19-enero-2026
Consulta	"capital humano" AND "desarrollo economico"[a]
Tamaño del corpus recuperado	6,688
Idioma seleccionado	Español
Tamaño del corpus final tras selección de idioma y conversión bibliométrica	6,055
Registros excluidos en la depuración lingüística y conversión	633 (9.5%)

[a]La ecuación de búsqueda se seleccionó como caso ilustrativo por tratarse de un tema transversal en economía, con el propósito de demostrar el flujo de recuperación, conversión y análisis bibliométrico implementado por el paquete.

Posteriormente, sobre la base bibliométrica resultante, se aplicó la función frequency_keywords() para estimar y visualizar la distribución de frecuencia de las palabras clave de autor (campo *DE*). El análisis se ejecutó de la siguiente manera:

```
frequency_keywords(df.redalyc, save = TRUE)
```

La Figura 2 presenta la distribución de frecuencia de las palabras clave más recurrentes. Los resultados muestran que, además de los términos directamente asociados a la consulta principal, emergen conceptos relacionados con educación, productividad, crecimiento económico y desarrollo social, lo que sugiere que la literatura que vincula capital humano y desarrollo económico se articula alrededor de un conjunto relativamente estable de ejes temáticos. Este tipo de resultado confirma la utilidad del análisis automatizado de frecuencias como primer nivel de exploración del corpus, permitiendo identificar rápidamente los conceptos dominantes y orientar análisis posteriores de mayor complejidad.

Desde una perspectiva metodológica, este resultado ilustra cómo redalycR facilita la generación de salidas reproducibles tanto gráficas como tabulares, exportables en formatos estándar, lo que permite su integración directa en procesos de redacción académica y análisis comparativo.

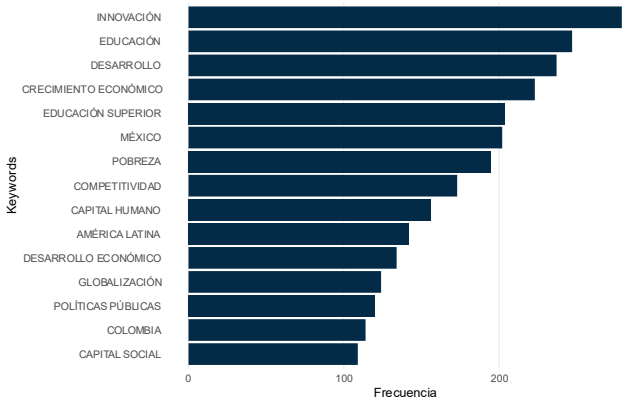

Fig. 2 Output de frequency_keywords.

Como segundo nivel de análisis, se construyó una red de co-ocurrencia de palabras clave mediante la función redalyc_keyword_networkPlot():

```
redalyc_keyword_networkPlot(df.redalyc, normalize = association",
                            cluster = "walktrap",
                            layout_type = "fruchterman",
                            min_freq = 2,
                            n = 80, save = TRUE)
```

Este enfoque permite examinar no solo la frecuencia individual de los términos, sino también sus relaciones estructurales dentro del corpus, capturando patrones de asociación y posibles agrupamientos temáticos. La Figura 3 muestra la red resultante, donde los nodos representan palabras clave y las aristas indican co-ocurrencia dentro de un mismo artículo. El tamaño de los nodos es proporcional a su grado de conexión, mientras que el peso de las aristas refleja la intensidad de la relación entre términos, normalizada mediante el índice de association. Adicionalmente, se aplicó un algoritmo de detección de comunidades para identificar subconjuntos de términos densamente interconectados.

Los resultados evidencian la existencia de clusters temáticos diferenciados, lo que sugiere que la literatura sobre capital humano y desarrollo económico no constituye un bloque homogéneo, sino que se organiza en sublíneas de investigación. Estas comunidades suelen agrupar conceptos vinculados, por ejemplo, a educación y formación, mercado laboral, productividad y políticas públicas, mostrando cómo los enfoques teóricos y empíricos se estructuran alrededor de combinaciones recurrentes de términos.

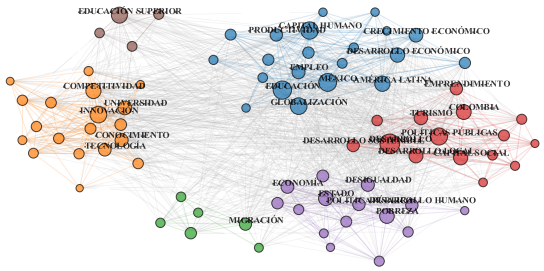

Fig. 3 Output de redalyc_keyword_networkPlot.

Este tipo de representación permite pasar de una lectura meramente descriptiva a una interpretación estructural del campo, facilitando la identificación de núcleos conceptuales y relaciones transversales entre temas. Desde la perspectiva del diseño del paquete, este resultado demuestra la capacidad de redalycR para reproducir análisis de redes bibliométricas comparables a los generados por herramientas consolidadas, manteniendo al mismo tiempo flexibilidad en la parametrización y control explícito de los supuestos metodológicos.

En conjunto, los resultados obtenidos muestran que redalycR permite operacionalizar, de forma reproducible y transparente, un flujo completo de análisis bibliométrico basado exclusivamente en datos abiertos provenientes de Redalyc. Estos resultados no buscan caracterizar exhaustivamente el campo analizado, sino demostrar la aplicabilidad del paquete como herramienta metodológica y de software científico, alineada con prácticas de reproducibilidad y extensibilidad.

IV. Discusión

Desde una perspectiva metodológica y de ingeniería de software científico, los resultados de este estudio evidencian que por medio de redalycR es posible estructurar un flujo reproducible de análisis bibliométrico utilizando exclusivamente datos abiertos provenientes de Redalyc, integrados en el ecosistema analítico de R.

Desde el punto de vista del diseño conceptual, redalycR se inspira claramente en paquetes consolidados como bibliometrix [36], que han demostrado el potencial del lenguaje R para el análisis bibliométrico y el mapeo científico. En ese sentido bibliometrix constituye un referente metodológico y técnico que influyó en la arquitectura de redalycR, particularmente en la forma de estructurar los datos bibliográficos y en la incorporación de análisis de palabras clave y redes de co-ocurrencia.

De forma práctica una contribución relevante de redalycR es su alineación con la producción científica en español y en América Latina, un aspecto que suele quedar subrepresentado en herramientas bibliométricas dominadas por bases de datos especializadas con Scopus, Web of Science, PubMed, entre otras. La posibilidad de acceder, estructurar y analizar de forma programática literatura científica en español, utilizando un flujo reproducible y abierto, representa un avance metodológico para investigadores, estudiantes de posgrado e instituciones que trabajan en contextos donde el acceso a bases de datos científicas es limitado. Más que sustituir enfoques existentes, redalycR amplía el abanico de herramientas disponibles y fortalece la capacidad analítica sobre un corpus que, en muchos casos, ha sido menos explorado desde la bibliometría computacional.

Asimismo, el desarrollo del paquete se inserta de manera natural en la lógica de ciencia abierta promovida por Redalyc, que ha avanzado en la provisión de mecanismos de acceso interoperables a sus metadatos y contenidos. En este marco, redalycR puede entenderse como un artefacto de software que operacionaliza dicha apertura, facilitando que los datos disponibles a través de APIs, OAI-PMH y formatos estructurados se integren efectivamente en flujos de investigación reproducible. El aporte del paquete no es normativo ni institucional, sino instrumental: traduce la disponibilidad de datos abiertos en capacidad analítica concreta para la comunidad académica.

No obstante, los resultados también ponen de manifiesto limitaciones actuales del paquete, que deben ser interpretadas como parte natural de una primera etapa de desarrollo. En particular, la ausencia de algunas variables bibliométricas presentes en bases de datos científicas como información completa de citaciones, referencias normalizadas o ciertos identificadores persistentes restringe el tipo de análisis que puede realizarse en esta versión. Estas limitaciones no solo definen el alcance actual del paquete, sino que abren una agenda clara de desarrollo futuro, tanto desde el punto de vista técnico como institucional. La ampliación de campos disponibles en Redalyc, ya sea mediante mayor especialización de los metadatos o a través de acuerdos de colaboración, podría potenciar significativamente los análisis posibles y fortalecer la interoperabilidad con otros flujos bibliométricos.

En términos de desarrollo futuro, una línea particularmente relevante es la extensión del análisis textual hacia los resúmenes (abstracts) y otros campos de contenido, lo que permitiría incorporar técnicas más avanzadas de minería de texto, modelado de tópicos y análisis semántico. La arquitectura modular de redalycR fue diseñada precisamente para facilitar este tipo de extensiones, manteniendo la trazabilidad del proceso y la coherencia metodológica. En este sentido, el paquete no debe interpretarse como un producto cerrado, sino como una plataforma en evolución, susceptible de incorporar nuevas funciones y capacidades conforme madure su desarrollo.

V. Conclusiones

Este trabajo presentó redalycR, un paquete en R orientado a la automatización del análisis bibliométrico mediante una arquitectura reproducible basada exclusivamente en datos abiertos provenientes de Redalyc. El estudio demostró que es posible estructurar un flujo completo desde la recuperación de literatura científica hasta el análisis descriptivo y relacional de palabras clave utilizando herramientas de software abierto y prácticas de investigación reproducible.

Desde una perspectiva metodológica, los resultados evidencian que la paquetización de procedimientos analíticos previamente implementados como scripts individuales contribuye a mejorar la estandarización, reutilización y transparencia de los análisis bibliométricos. En este sentido, redalycR se posiciona como un complemento a herramientas consolidadas del ecosistema R, inspirado en enfoques existentes pero adaptado a las particularidades y limitaciones de una base de datos abierta no propietaria.

Asimismo, el paquete aporta una vía concreta para potenciar el análisis bibliométrico de literatura científica en español, fortaleciendo la capacidad analítica sobre corpus que suelen estar subrepresentados en estudios basados en bases de datos comerciales. Al operacionalizar el acceso abierto promovido por Redalyc, redalycR facilita que investigadores y estudiantes integren datos abiertos en flujos computacionales replicables, sin requerir infraestructuras cerradas o licencias propietarias.

Finalmente, este trabajo debe entenderse como un primer paso dentro de una línea de desarrollo en curso. Las limitaciones actuales del paquete particularmente en relación con la disponibilidad de variables bibliométricas y el análisis de contenido textual más profundo delinean una agenda clara de trabajo futuro, orientada a la incorporación de nuevas funciones y al fortalecimiento del ecosistema de análisis bibliométrico basado en ciencia abierta. En conjunto, redalycR contribuye al avance de prácticas metodológicas reproducibles y sienta las bases para desarrollos posteriores tanto a nivel de software como de investigación académica.

REFERENCES

[1] D. N. Boote and P. Beile, "Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation," *Educational Researcher*, vol. 34, no. 6, pp. 3–15, 2005, doi: 10.3102/0013189X034006003.

[2] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data: A Survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, 2017, doi: 10.1109/TBDATA.2016.2641460.

[3] J. Portenoy and J. D. West, "Constructing and evaluating automated literature review systems," *Scientometrics*, vol. 125, no. 3, pp. 3233–3251, 2020, doi: 10.1007/s11192-020-03490-w.

[4] L. Siguenza-Guzman, V. Saquicela, E. Avila-Ordóñez, J. Vandewalle, and D. Cattrysse, "Literature Review of Data Mining Applications in Academic Libraries," *The Journal of Academic Librarianship*, vol. 41, no. 4, pp. 499–510, 2015, doi: 10.1016/j.acalib.2015.06.007.

[5] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R," *J. Stat. Soft.*, vol. 25, no. 5, 2008, doi: 10.18637/jss.v025.i05.

[6] A. Żarczyńska, "Nicola De Bellis: Bibliometrics And Citation Analysis, from the Science Citation Index to Cybermetrics, Lanham, Toronto, Plymouth 2009," *TSB*, vol. 5, 1 (8), 2012, doi: 10.12775/TSB.2012.009.

[7] G. Mathew, A. Agrawal, and T. Menzies, "Finding Trends in Software Research," *IIEEE Trans. Software Eng.*, vol. 49, no. 4, pp. 1397–1410, 2023, doi: 10.1109/TSE.2018.2870388.

[8] V. Garousi and M. V. Mäntylä, "Citations, research topics and active countries in software engineering: A bibliometrics study," *Computer Science Review*, vol. 19, pp. 56–77, 2016, doi: 10.1016/j.cosrev.2015.12.002.

[9] P. Raulamo-Jurvanen, M. V. Mantyla, and V. Garousi, "Citation and Topic Analysis of the ESEM Papers," in *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Beijing, China, 2015, pp. 1–4.

[10] A. Jadhav, M. Kaur, and F. Akter, "Evolution of Software Development Effort and Cost Estimation Techniques: Five Decades Study Using Automated Text Mining Approach," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–17, 2022, doi: 10.1155/2022/5782587.

[11] S. Matysik, J. Wiśniewska, and P. K. Frankowski, "EmbedSLR: an open-source python framework for efficient embedding-based screening and bibliometric validation in systematic literature review," *SoftwareX*, vol. 32, p. 102416, 2025, doi: 10.1016/j.softx.2025.102416.

[12] M. Abbasi, P. Váz, J. Silva, F. Cardoso, F. Sá, and P. Martins, "Comparative Performance Analysis of Large Language Models for Structured Data Processing: An Evaluation Framework Applied to Bibliometric Analysis," *Applied Sciences*, vol. 16, no. 2, p. 669, 2026, doi: 10.3390/app16020669.

[13] D. Mitroulias and S. Sioutas, "A Systematic Review and Bibliometric Analysis of Automated Multiple-Choice Question Generation," *BDCC*, vol. 10, no. 1, p. 35, 2026, doi: 10.3390/bdcc10010035.

[14] J. Zhou, H. Zhang, and L. Ding, "The Global Research Landscape of AI in education: A Bibliometric analysis of Pathway Evolution and Frontier Issues," in *2025 11th International Conference on Education and Training Technologies (ICETT)*, Macao, China, 2025, pp. 90–98.

[15] M. Maričić, A. Dacić-Pilčević, V. Jeremić, and V. Uskoković, "Navigating the landscape," *Zb. rad. Ekon. fak. Rij. (Online)*, vol. 42, no. 2, pp. 479–507, 2024, doi: 10.18045/zbefri.2024.2.5.

[16] M. R. C. Santos, "Blended Learning Compared to Online Learning in Business, Management, and Accounting Education," in *Advances in Logistics, Operations, and Management Science, Interdisciplinary and Practical Approaches to Managerial Education and Training*, J. Wang, L. C. Carvalho, N. Teixeira, and P. Pardal, Eds.: IGI Global, 2022, pp. 46–55.

[17] H. Osorto, "Optimizing academic literature review using Textmining in R: An automated approach," in *Proceedings of the 23rd LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI): "Engineering, Artificial Intelligence, and Sustainable Technologies in service of society"*, Jul. 2025.

[18] T. Warin, "Global Research on Coronaviruses: An R Package," *Journal of medical Internet research*, vol. 22, no. 8, e19615, 2020, doi: 10.2196/19615.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria. [Online]. Available: https://www.r-project.org/

[20] Hadley Wickham, *httr2: Perform HTTP Requests and Process the Responses.* [Online]. Available: https://cran.r-project.org/package=httr2

[21] Jeroen Ooms, "The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects," *arXiv:1403.2805 [stat.CO]*, 2014.

[22] Kirill Müller and Hadley Wickham, *tibble: Simple Data Frames.* [Online]. Available: https://cran.r-project.org/package=tibble

[23] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan, *dplyr: A Grammar of Data Manipulation.* [Online]. Available: https://cran.r-project.org/package=dplyr

[24] Hadley Wickham, Davis Vaughan, and Maximilian Girlich, *tidyr: Tidy Messy Data.* [Online]. Available: https://cran.r-project.org/package=tidyr

[25] Hadley Wickham and Lionel Henry, *purrr: Functional Programming Tools.* [Online]. Available: https://cran.r-project.org/package=purrr

[26] Hadley Wickham, *stringr: Simple, Consistent Wrappers for Common String Operations.* [Online]. Available: https://cran.r-project.org/package=stringr

[27] Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org/

[28] Dirk Eddelbuettel, *digest: Create Compact Hash Digests of R Objects.* [Online]. Available: https://cran.r-project.org/package=digest

[29] Gábor Csárdi, *cli: Helpers for Developing Command Line Interfaces.* [Online]. Available: https://cran.r-project.org/package=cli

[30] G. Csárdi *et al., igraph for R: R interface of the igraph library for graph theory and network analysis*: Zenodo, 2025.

[31] Douglas Bates, Martin Maechler, and Mikael Jagan, *Matrix: Sparse and Dense Matrix Classes and Methods.* [Online]. Available: https://cran.r-project.org/package=Matrix

[32] Jeroen Ooms, *cld2: Google's Compact Language Detector 2.* [Online]. Available: https://cran.r-project.org/package=cld2

[33] Jeroen Ooms, *cld3: Google's Compact Language Detector 3.* [Online]. Available: https://cran.r-project.org/package=cld3

[34] Jeroen Ooms, *writexl: Export Data Frames to Excel 'xlsx' Format.* [Online]. Available: https://cran.r-project.org/package=writexl

[35] Hadley Wickham, Lionel Henry, Thomas Lin Pedersen, T Jake Luciani, Matthieu Decorde, and Vaudor Lise, *svglite: An 'SVG' Graphics Device.* [Online]. Available: https://cran.r-project.org/package=svglite

[36] M. Aria and C. Cuccurullo, "bibliometrix : An R-tool for comprehensive science mapping analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, 2017, doi: 10.1016/j.joi.2017.08.007.