Chapter 4 / Capítulo 4

Applied bibliometrics. From data to publication (Spanish Edition)
ISBN: 978-9915-9680-6-3
DOI: 10.62486/978-9915-9680-6-3.ch04

©2025 The authors. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

DATA COLLECTION / OBTENCIÓN DE DATOS

Del Caos al Conocimiento Organizado

4.1. Estrategias de búsqueda para estudios bibliométricos

En el ámbito de la bibliometría, la calidad de los resultados depende fundamentalmente de una estrategia de búsqueda rigurosa y bien planificada. Una aproximación deficiente en esta fase inicial puede comprometer todo el estudio, generando sesgos significativos, omitiendo literatura fundamental o incorporando material irrelevante que distorsiona los análisis. Las consecuencias de un diseño inadecuado son particularmente graves en este campo, donde la integridad de los hallazgos está directamente vinculada a la exhaustividad y precisión del corpus documental recuperado.

Este capítulo ofrece un marco metodológico sistemático para desarrollar cadenas de búsqueda óptimas en las principales plataformas bibliográficas (Web of Science, Scopus, PubMed, entre otras). El enfoque presentado combina principios teóricos con aplicaciones prácticas:

- Técnicas avanzadas para la formulación de ecuaciones de búsqueda balanceadas
- Estrategias de validación para asegurar cobertura completa sin sacrificar precisión
- Ejemplos concretos extraídos de diversas disciplinas científicas
- Errores frecuentes y cómo evitarlos, basados en análisis de casos reales

El método propuesto enfatiza la necesidad de adaptar cada búsqueda a las particularidades del sistema documental de cada campo de estudio, considerando factores como: la terminología disciplinar específica, las características de indización de cada base de datos, y los sesgos inherentes a los diferentes sistemas de recuperación de información.

Mediante este enfoque estructurado, los investigadores podrán construir conjuntos documentales robustos que sirvan como base fiable para análisis bibliométricos avanzados, desde estudios de mapeo científico hasta evaluaciones de impacto. La metodología presentada ha sido validada empíricamente en múltiples proyectos de investigación, demostrando su eficacia para producir resultados representativos y replicables en diversas áreas del conocimiento.

4.1.1. Diseño de la estrategia de búsqueda

Definir el alcance: la base de una búsqueda bibliométrica efectiva

El primer paso crítico en cualquier estudio bibliométrico es establecer con precisión los límites y objetivos de la investigación. Este proceso de delimitación requiere responder preguntas fundamentales que determinarán todo el diseño metodológico posterior.

La pregunta de investigación central debe formularse con claridad, ya que definirá el tipo de análisis a realizar. ¿El objetivo es identificar tendencias temporales en un campo? ¿Evaluar el impacto de ciertas instituciones? ¿Mapear redes de colaboración internacional? Cada uno de estos propósitos exigirá estrategias de búsqueda diferentes.

Los criterios de inclusión constituyen los filtros esenciales para garantizar relevancia y manejabilidad de los resultados:

- El período temporal debe justificarse según la evolución del campo (ej: en áreas emergentes, los últimos 5 años pueden ser suficientes; en disciplinas consolidadas, puede necesitarse una década o más).
- Los tipos documentales seleccionados (artículos, revisiones, patentes, tesis) deben alinearse con las prácticas de comunicación de cada disciplina.

- La decisión sobre idiomas implica un equilibrio entre exhaustividad (incluir contribuciones en múltiples lenguas) y factibilidad (el predominio del inglés en muchas áreas científicas).

Un ejemplo aplicado, por ejemplo, para estudiar "*La producción científica sobre inteligencia artificial aplicada a medicina entre 2015-2025*", una delimitación adecuada podría especificar:

- Solo artículos completos en revistas con revisión por pares
- Documentos indexados en las principales bases de datos (WoS, Scopus)
- Publicaciones en inglés (dado su predominio en estas disciplinas técnicas)
- Exclusión de patentes y literatura gris para mantener el foco en investigación básica

Esta fase de planificación, aunque a menudo subestimada, es determinante. Una gran parte de los errores en bibliometrías se originan en una definición imprecisa del alcance inicial. Una delimitación clara evita posteriormente la inclusión de ruido documental o la omisión de literatura clave, problemas que pueden comprometer la validez de los hallazgos.

Seleccionar bases de datos

Teniendo en cuenta el alcance planificado se debe seleccionar la base de datos sobre la cual se hará la búsqueda, por lo que son necesarios conocimientos básicos sobre el contenido y bondades de cada base de datos. Cada base tiene ventajas y sesgos:

Tabla 4.1. Características de las principales bases de datos científicas

Base de Datos	Cobertura	Filtros Útiles	Limitaciones
Scopus	+25,000 revistas, fuerte en STEM	Análisis de citas integrado	Costoso, débil en humanidades
Web of Science	+20,000 revistas, núcleo histórico	Índice h incorporado	Sesgo hacia EE.UU./ Europa
PubMed	+30M registros en biomedicina	Filtros por tipo de estudio (ensayos clínicos)	Sin datos de citas completos
Google Scholar	Libre, incluye libros y grises	Búsqueda semántica	Sin herramientas de análisis
Dimensions	Integra publicaciones y patentes	Datos de financiamiento	Cobertura desigual por región

Es oportuno explicar que, para evitar los sesgos/limitaciones de las bases de datos, pueden pueden emplearse varias en un mismo estudio (ej: Scopus + PubMed).

4.1.2. Construcción de cadenas de búsqueda

Una vez definido con claridad cuál es el tema sobre el cual se hará la búsqueda se deben seleccionar las palabras más representativas de la investigación.

Palabras clave y operadores booleanos

La efectividad de una búsqueda de información para una investigación bibliométrica depende, en gran medida, de la correcta selección de los términos clave y su combinación mediante operadores lógicos. El operador AND restringe los resultados, conectando conceptos que deben aparecer simultáneamente (ejemplo: «*machine learning*» AND «*healthcare*» recupera solo documentos que mencionen ambos temas). Por el contrario, OR amplía el alcance al incluir sinónimos o variantes conceptuales (ejemplo: «*deep learning*» OR «*neural networks*»). Para excluir temas no relevantes, se emplea NOT, aunque debe usarse con precaución para

no eliminar literatura útil (ejemplo: «*cancer*» NOT «*breast cancer*» podría omitir estudios transversales importantes).

Las comillas garantizan búsquedas exactas de frases completas, evitando resultados dispersos («climate change» es más preciso que climate AND change). Los comodines capturan variantes morfológicas de una raíz común (ejemplo: educat incluye «education», «educative», «educational»).

Ejemplo de cadena optimizada:
(«artificial intelligence» OR «AI») AND («medical diagnosis» OR «clinical decision») NOT «robotics»

Esta consulta recupera documentos sobre IA aplicada a diagnóstico médico, excluyendo aquellos centrados en robótica.

Los operadores anteriores son comunes para la mayoría de bases de datos y buscadores, pero hay algunas bases de datos que incorporan nuevos en aras de ganar en especificidad de la búsqueda. De todas formas la lógica que siguen estos operadores es la misma independientemente de donde se implementen.

Filtros avanzados para precisión

Muchas veces solo con lo anterior no basta para recuperar el contenido requerido o trazado, o abarca mucho más de lo que se planeó. Las bases de datos permiten afinar búsquedas mediante filtros específicos:

- Por campo: En Scopus, **TITLE-ABS-KEY ()** busca en título, resumen y palabras clave; en PubMed, **TI/AB** cubre título y resumen. Limitar la búsqueda solo al título (**TITLE()**) subestima la literatura relevante, ya que muchos conceptos clave pueden aparecer únicamente en el resumen.
- Por impacto: Filtros como **CITES > 50** (WoS) identifican trabajos altamente citados, útiles para estudios de influencia científica.
- Por institución: **AFFIL("Harvard University")** localiza la producción de una universidad específica, esencial para análisis de colaboraciones.

Error frecuente y solución:

Problema: Usar solo TITLE() puede perder documentos relevantes
Solución: Combinar siempre título, resumen y palabras clave (TITLE-ABS-KEY() en Scopus o TI/AB/KW en WoS).

4.1.3. Técnicas específicas por base de datos

En Scopus, la sintaxis avanzada permite filtrar con precisión mediante operadores como TITLE-ABS-KEY(), que abarca título, resumen y palabras clave simultáneamente. Por ejemplo, una consulta como TITLE-ABS-KEY(«blockchain» AND «supply chain») AND PUBYEAR > 2015 recupera exclusivamente publicaciones recientes sobre aplicaciones de blockchain en cadena de suministro.

Web of Science ofrece funcionalidades únicas para el análisis de redes institucionales. Su sintaxis TS=() para términos de búsqueda combinada con AD=() para afiliaciones permite mapear colaboraciones entre centros de investigación específicos. Una consulta como TS=(«quantum computing») AND AD=(«MIT» OR «Stanford») identifica la producción conjunta de estas universidades en computación cuántica.

Para investigaciones biomédicas, PubMed se distingue por su sistema de vocabulario controlado MeSH (Medical Subject Headings). Este tesauro jerárquico resuelve problemas de sinonimia y ambigüedad terminológica. Una búsqueda como «COVID-19»[Mesh] AND «Vaccines»[Mesh] AND «clinical trials»[Publication Type] garantiza cobertura completa sobre ensayos clínicos de vacunas contra COVID-19, evitando los sesgos de las búsquedas por texto libre. La precisión de PubMed lo hace indispensable para revisiones sistemáticas en ciencias de la salud.

Cada plataforma exige estrategias específicas para optimizar precisión y recall. Scopus y WoS priorizan control terminológico y capacidades analíticas integradas, mientras PubMed ofrece vocabulario estandarizado y Google Scholar amplitud documental a costa de transparencia metodológica. La elección entre ellas debe basarse en los objetivos del estudio: análisis de citación y colaboración (WoS/Scopus), revisiones clínicas rigurosas (PubMed) o identificación de literatura no convencional (Google Scholar). Un diseño de investigación robusto frecuentemente combina múltiples bases para compensar sus respectivas limitaciones.

Recientemente, herramientas como Dimensions han emergido como alternativas prometedoras, integrando publicaciones, patentes y datasets en un solo entorno con API abierta. Su sintaxis flexible (ej: search publications for «CRISPR» where year > 2020) y cobertura multidisciplinar están redefiniendo los estándares para búsquedas bibliométricas complejas. Sin embargo, la adopción generalizada de estos nuevos sistemas aún enfrenta desafíos de interoperabilidad con las métricas tradicionales.

En resumen, los operadores y filtros específicos por bases de datos son:

Scopus

Operadores booleanos y de búsqueda avanzada:

- AND, OR, AND NOT → Combinan términos.
- → Truncamiento (ej.: comput busca “computer”, “computing”, etc.).
- “ “ → Búsqueda exacta (ej.: “machine learning”).
- W/n → Palabras dentro de n términos de distancia (ej.: AI W/3 healthcare).
- PRE/n → Primera palabra antes de la segunda, con n términos de distancia (ej.: digital PRE/2 transformation).

Filtros avanzados:

- Año: PUBYEAR > 2020
- Tipo de documento: DOCTYPE(ar) (artículos) o DOCTYPE(re) (revisiones).
- Campo específico:
  - TITLE-ABS-KEY(“blockchain”) → Busca en título, resumen o palabras clave.
  - AFFIL(“Harvard University”) → Por afiliación.
- Índices de citas: REFERENCES() (ej.: REFERENCES(123456789)).

Ejemplo combinado: TITLE-ABS-KEY(“quantum computing”) AND PUBYEAR > 2018 AND AFFIL(“MIT”) AND DOCTYPE(ar)

Web of Science (WoS)

Operadores y sintaxis:

- AND, OR, NOT → Operadores estándar.
- → Truncamiento (ej.: bio para “biology”, “biotechnology”).
- “ “ → Frase exacta.

- NEAR/x → Términos cercanos (ej.: climate NEAR/2 change).
- SAME → Términos en el mismo campo (útil para direcciones: UNIV COLORADO SAME BOULDER).

Filtros avanzados:

- Categorías WC: WC=("Computer Science").
- Año: PY=2020-2023.
- Tipo de documento: DT=("Article").
- Búsqueda por campo:
  - TS=("neural network") → Tema (título, resumen, palabras clave).
  - AU=("Smith J") → Autor.
  - OG=("Stanford University") → Organización.

Ejemplo combinado: TS=("CRISPR" AND "gene editing") AND PY=(2020-2023) AND WC=("Biotechnology") AND DT=("Review")

PubMed

Operadores y sintaxis:

- AND, OR, NOT → Operadores booleanos.
- → Truncamiento (ej.: immun para "immune", "immunity").
- " " → Frase exacta.
- [Field] → Búsqueda por campo (ej.: «COVID-19»[Title]).

Filtros avanzados (usando etiquetas):

- Tipo de artículo: "review"[Publication Type].
- Fecha: "2020/01/01"[Date - Publication] : "2023"[Date - Publication].
- Campos específicos:
  - "deep learning"[Title/Abstract].
  - "NIH"[Affiliation].
  - - "1AU"[Author] → Autor principal.
- - Filtros predefinidos:
  - "free full text"[Filter].
  - "clinical trial"[Filter].

Ejemplo combinado: "artificial intelligence"[Title/Abstract] AND "diagnosis"[Title]) AND "2020"[Date - Publication] AND "systematic review"[Publication Type]

4.1.4. Validación y documentación: garantizando rigor en la búsqueda bibliométrica

La fase de validación es fundamental para asegurar que los resultados de una búsqueda bibliométrica sean exhaustivos, precisos y reproducibles. Un proceso riguroso comienza con una prueba piloto, donde se analizan críticamente los primeros 50-100 registros obtenidos. Este muestreo inicial permite identificar problemas frecuentes: términos demasiado amplios que generan ruido documental o, por el contrario, cadenas de búsqueda demasiado restrictivas que omiten literatura relevante. La iteración es clave aquí, cada ajuste en los operadores booleanos o en la selección de términos debe probarse hasta alcanzar un equilibrio óptimo entre precisión y sensibilidad.

Para garantizar transparencia metodológica, el protocolo PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) ofrece un marco estandarizado, aplicable también a estudios bibliométricos. Su enfoque sistemático requiere documentar detalladamente todo

el proceso de selección: el número total de registros identificados en cada base de datos, los criterios de exclusión aplicados (duplicados, documentos fuera del alcance temporal o temático), y las decisiones tomadas en cada fase de filtrado. El diagrama de flujo PRISMA se convierte así en una herramienta visual poderosa que resume el proceso de depuración del corpus documental, permitiendo a otros investigadores evaluar la solidez metodológica del estudio.[1]

La documentación detallada debe incluir, además: la sintaxis exacta de todas las cadenas de búsqueda utilizadas (incluyendo variantes descartadas), las fechas de ejecución de las consultas (importante dado el carácter dinámico de las bases de datos), y cualquier limitación conocida en la cobertura de las fuentes utilizadas. Tener claridad de todo esto es crucial desde la fase inicial de la investigación pues como se verá en capítulos siguientes todo esto debe quedar detallado en el informe final.

Los investigadores experimentados suelen complementar este proceso con validación externa, consultando a expertos temáticos para verificar que la búsqueda no haya omitido contribuciones seminales, o comparando sus resultados con revisiones sistemáticas existentes en el área. Este paso adicional ayuda a detectar posibles sesgos en la estrategia de búsqueda que podrían afectar las conclusiones del estudio. La combinación de validación interna (mediante protocolos estandarizados) y externa (a través de juicio experto) constituye la mejor garantía para producir resultados bibliométricos robustos y confiables.

Este enfoque metódico transforma la búsqueda bibliográfica de un mero procedimiento técnico en un proceso investigativo riguroso, donde cada decisión está justificada y documentada. La inversión de tiempo en esta fase crítica se traduce directamente en la calidad y credibilidad de los hallazgos posteriores, evitando los sesgos y limitaciones que frecuentemente afectan a los estudios bibliométricos realizados sin esta sistematicidad.

Esta reducción progresiva debe documentarse meticulosamente, especificando los motivos de exclusión en cada etapa. La trazabilidad completa del proceso no solo fortalece la validez interna del estudio, sino que facilita su replicación o actualización futura.

4.1.5. Errores comunes y soluciones

Una búsqueda demasiado amplia, que arroja más de 10 000 registros, suele ser inmanejable para análisis bibliométricos detallados y puede incluir gran cantidad de ruido documental. Este problema ocurre frecuentemente al utilizar términos genéricos sin filtros adecuados. La solución pasa por incorporar restricciones estratégicas como límites temporales (ejemplo: últimos 10 años), filtrado por tipo de documento (solo artículos de investigación) o enfoque en campos específicos (título y resumen en lugar de texto completo).

En el extremo opuesto, una búsqueda demasiado estrecha puede dejar fuera contribuciones seminales por usar terminología muy específica. Esto es particularmente problemático en campos emergentes donde los conceptos aún no han estandarizado su nomenclatura. Para evitarlo, es esencial construir cadenas de búsqueda que incorporen sinónimos y variantes terminológicas mediante operadores OR, además de consultar tesauros disciplinares como MeSH en PubMed o los términos controlados de Scopus.

Un error frecuente en ciertas disciplinas es ignorar libros y actas de conferencias, lo que genera un sesgo sistemático contra áreas como humanidades o ciencias computacionales respectivamente. Las bases tradicionales (WoS/Scopus) tienen cobertura limitada en estos

formatos, por lo que debe complementarse la búsqueda con Google Scholar para libros o Dimensions para proceedings. En estudios interdisciplinares, esta omisión puede significar perder gran parte de la literatura relevante.

La falta de documentación de la estrategia de búsqueda compromete la reproducibilidad del estudio, un principio fundamental en investigación. La solución implica guardar sistemáticamente no solo la cadena final, sino todas sus iteraciones, especificando: plataforma utilizada, fecha de búsqueda, filtros aplicados y número de resultados obtenidos en cada paso. Formatos estándar como los diagramas PRISMA o plantillas de protocolos de revisión sistemática pueden adaptarse para este fin. Herramientas como Zotero o Mendeley permiten exportar los registros de búsqueda con metadatos completos para auditoría posterior.

Estos problemas y sus soluciones ilustran cómo el diseño de búsquedas bibliométricas es un proceso iterativo que requiere equilibrio entre exhaustividad y precisión. La experiencia muestra que dedicar el 30 % del tiempo total del estudio a perfeccionar y validar la estrategia de búsqueda puede mejorar significativamente la calidad de los resultados finales. La incorporación de revisiones por pares expertos en la fase de diseño metodológico ayuda a identificar potenciales omisiones o sesgos antes de ejecutar las búsquedas definitivas.

El uso creciente de técnicas de *machine learning* para el *screening* automatizado de grandes volúmenes de resultados está transformando este proceso, pero la construcción inicial de cadenas de búsqueda precisas sigue siendo fundamental. Plataformas como ASReview o Rayyan pueden luego ayudar en la fase de selección, pero dependen críticamente de una estrategia de recuperación documental bien diseñada desde el inicio. Esta sinergia entre métodos tradicionales y nuevas tecnologías marca el camino hacia búsquedas bibliométricas más eficientes y completas.

4.2 Herramientas alternativas para búsqueda bibliométrica: ampliando horizontes

Aunque Scopus y Web of Science siguen siendo referentes en análisis bibliométrico, existen alternativas valiosas que democratizan el acceso a la información científica. Google Scholar, con su amplia cobertura de literatura gris, junto con herramientas como *Publish or Perish* y gestores de referencias avanzados, ofrecen posibilidades únicas para investigadores con recursos limitados o que trabajan con formatos no convencionales

4.2.1 Google Scholar: ventajas y limitaciones prácticas

La principal fortaleza de Google Scholar radica en su capacidad para indexar contenido que frecuentemente queda fuera de las bases comerciales: desde *preprints* en arXiv hasta tesis doctorales y documentos de trabajo. Esta exhaustividad lo hace particularmente útil para estudios en humanidades o disciplinas emergentes donde la publicación formal puede ser más lenta. Sin embargo, su uso sistemático presenta desafíos técnicos significativos. El límite aproximado de 100 consultas diarias antes del bloqueo IP obliga a estrategias de muestreo cuidadosas, mientras que la falta de metadatos estructurados complica los análisis bibliométricos avanzados.

Para investigadores con habilidades técnicas, la librería “scholarly” de Python ofrece un balance entre automatización y respeto a los términos de servicio. El proceso básico implica: instalación mediante pip, configuración de consultas específicas, y extracción controlada de datos. Un *script* podría incluir:

```python
import csv
from scholarly import scholarly
```

```
search_query = scholarly.search_pubs("blockchain in healthcare")
with open('resultados.csv', 'w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerow(["Título", "Autores", "Año", "Citas"])
    for i in range(50):   Límite responsable
        try:
            pub = next(search_query)
            writer.writerow([
                pub['bib'].get('title', ''),
                pub['bib'].get('author', ''),
                pub['bib'].get('year', ''),
                pub.get('num_citations', 0)
            ])
        except StopIteration:
            break
```

Este enfoque incluye salvaguardas clave: límite de registros, manejo de errores, y almacenamiento en formato CSV para análisis posterior. Para usuarios sin programación, extensiones como Scholarcy ofrecen funcionalidad básica de extracción directamente desde el navegador, aunque con menos personalización.

El *scraping* de datos académicos debe seguir principios claros: intervalos entre consultas que simulen comportamiento humano, respeto a los robots.txt de cada sitio, y uso exclusivo para fines de investigación. Es crucial recordar que muchos documentos en Google Scholar están protegidos por derechos de autor, por lo que la extracción masiva de textos completos sin permiso constituye violación ética. Las mejores prácticas recomiendan:

- Limitar las extracciones a metadatos básicos (título, autores, citas)
- Almacenar solo la información necesaria para el análisis
- Citar adecuadamente todas las fuentes recuperadas
- Verificar periódicamente los términos de servicio de las plataformas

Estas herramientas alternativas funcionan mejor como complemento, no reemplazo, de las bases tradicionales. Un flujo de trabajo robusto podría comenzar con búsquedas sistemáticas en Scopus/WoS, seguido de un barrido en Google Scholar para capturar literatura adicional, finalizando con gestores como Zotero o Mendeley para organización y de duplicación. Esta triangulación metodológica mitiga los sesgos inherentes de cada plataforma mientras maximiza la cobertura documental.

4.2.2 Harzing's Publish or Perish

Publish or Perish (PoP) no solo extrae datos de Google Scholar, sino que también ofrece integración con bases de datos académicas clave como Scopus, Web of Science (WoS) y PubMed, ampliando significativamente su utilidad para investigadores que requieren análisis más estructurados. Esta capacidad multiplataforma lo convierte en una solución versátil para estudios bibliométricos comparativos, permitiendo contrastar métricas de impacto entre diferentes fuentes y detectar posibles discrepancias en la cobertura de citaciones. En el capítulo 3 se abordó en detalle como instalar esta herramienta así como sus ventajas y particularidades. De manera general, para hacer una búsqueda con ella se procede de la siguiente manera.

1. Descarga e instalación:
    - Disponible en https://harzing.com/resources/publish-or-perish (Windows/Mac/Linux).

2. Primera búsqueda:
    - Abre el software y haz clic en “New Query”.
    - Ingresa términos de búsqueda (ej: `”artificial intelligence education”`).

4. Exportar resultados:
    - Ve a File > Save As y elige CSV/Excel para análisis en Bibliometrix o Excel.

4.2.3 Gestores de referencias para análisis bibliométrico avanzado

Los gestores de referencias como Zotero y Mendeley han evolucionado más allá de simples organizadores bibliográficos, convirtiéndose en herramientas poderosas para la extracción y procesamiento masivo de datos bibliométricos. Su capacidad para capturar metadatos estructurados directamente desde portales académicos o archivos PDF los hace indispensables en flujos de trabajo de investigación cuantitativa.

Zotero destaca por su integración perfecta con bases de datos como Scopus y Web of Science. Al utilizar su conector de navegador, los investigadores pueden importar cientos de referencias con un solo clic, preservando información crítica como DOI, afiliaciones institucionales y campos de citación. El proceso de exportación a CSV (a través del menú contextual) genera tablas listas para su análisis en software especializado, aunque requiere atención durante la selección de campos personalizados para asegurar que todos los metadatos relevantes sean incluidos.

Mendeley ofrece ventajas distintas, particularmente en el manejo de colecciones de PDFs existentes. Su motor de reconocimiento de documentos puede extraer metadatos incluso de archivos sin información embebida, aunque con precisión variable. La función “***Verify Document Details***” permite corregir errores comúnmente encontrados en artículos más antiguos o de revistas menos establecidas. Para proyectos bibliométricos, su capacidad de exportación en formato RIS resulta especialmente valiosa, ya que este estándar conserva relaciones entre documentos que formatos tabulares como CSV podrían perder.

Un desafío recurrente con ambas herramientas es la calidad inconsistente de metadatos. Estudios recientes muestran que hasta el 15 % de los artículos en gestores de referencias requieren corrección manual de campos básicos como año de publicación o lista completa de autores. Aquí, APIs como CrossRef se vuelven complementos esenciales. El ejemplo en Python demostrado (utilizando la librería habanero) ilustra cómo automatizar la recuperación de metadatos faltantes mediante consultas basadas en títulos o fragmentos de texto. Este enfoque híbrido -combinando extracción automática con verificación programática- puede mejorar significativamente la eficiencia en la preparación de datos bibliométricos.

Consideraciones técnicas avanzadas emergen al trabajar con grandes volúmenes de referencias. Zotero permite la ejecución de scripts personalizados a través de su API interna (Zotero API), posibilitando tareas como:

- Normalización automática de nombres institucionales
- Detección y fusión de duplicados basada en similitud semántica
- Extracción de citas entre documentos mediante análisis de referencias

Mientras tanto, Mendeley Desktop (a diferencia de su versión web) ofrece acceso directo a su

base de datos SQLite local, permitiendo consultas complejas que superan las limitaciones de la interfaz gráfica. Investigadores con conocimientos de SQL pueden así realizar transformaciones avanzadas de datos antes de la exportación final.

La elección entre estas herramientas depende frecuentemente del flujo de trabajo específico. Zotero brilla en proyectos colaborativos gracias a su sincronización en la nube y permisos granulares, mientras que Mendeley puede preferirse en entornos donde la gestión de PDFs masiva es prioritaria. Ambos, sin embargo, comparten el potencial de servir como puente entre la recolección inicial de literatura y el análisis bibliométrico sofisticado en plataformas como VOSviewer o Bibliometrix.

Estos gestores pueden utilizarse para extraer datos de PDFs o incluso de la web para posteriormente ser utilizados en los estudios bibliométricos.

4.3 Formatos de exportación en bibliometría: interoperabilidad y análisis

La selección del formato de exportación en bibliometría representa una decisión crítica que impacta directamente en la calidad y profundidad del análisis posterior. El formato RIS (*Research Information Systems*), desarrollado originalmente para bases de datos académicas, sobresale por su capacidad de preservar metadatos complejos como las relaciones de citación y las estructuras jerárquicas de autoría. Este formato mantiene un equilibrio único entre legibilidad humana y procesamiento automatizado, siendo particularmente valioso cuando se trabaja con herramientas de análisis de redes como VOSviewer. Su estructura de etiquetas estandarizadas (como AU para autores o PY para año de publicación) permite una transición fluida entre diferentes plataformas sin pérdida de información crítica.

BibTeX, por su parte, ofrece ventajas distintas en entornos donde prima la escritura académica técnica. Su sintaxis minimalista y su integración nativa con sistemas LaTeX lo hacen insustituible para investigadores que necesitan gestionar referencias mientras redactan. Sin embargo, esta misma simplicidad se convierte en limitación al abordar análisis bibliométricos avanzados, donde se requieren metadatos más ricos sobre citaciones, afiliaciones institucionales o indicadores de impacto. Herramientas modernas como PyBibX han extendido su utilidad al permitir cierto grado de procesamiento bibliométrico, pero aún requieren transformaciones adicionales para alcanzar el nivel de detalle que proporcionan otros formatos.

El formato CSV representa el puente más directo hacia el análisis cuantitativo avanzado. Su estructura tabular se adapta perfectamente a entornos estadísticos como R (a través de Bibliometrix) o Python (con Pandas), donde la manipulación de grandes volúmenes de datos exige flexibilidad. No obstante, esta misma adaptabilidad conlleva riesgos: las exportaciones automáticas desde algunas plataformas pueden aplanar estructuras de datos complejas, como listas de autores o jerarquías de instituciones, en cadenas de texto que requieren posterior procesamiento. La clave para aprovechar CSV en bibliometría está en garantizar que la exportación inicial capture todos los metadatos relevantes en columnas bien definidas.

4.3.1 Conversión entre formatos: herramientas y usos en software bibliométrico

La realidad del trabajo bibliométrico contemporáneo exige frecuentemente navegar entre múltiples formatos, cada uno optimizado para diferentes etapas del flujo de investigación. Las conversiones entre RIS, BibTeX y CSV no son meras traducciones técnicas, sino procesos que pueden ampliar o limitar las posibilidades analíticas. Herramientas como Zotero realizan estas transformaciones de manera aceptable para usos básicos, pero los proyectos de envergadura requieren enfoques más sofisticados.

En el ecosistema actual de software bibliométrico, cada herramienta principal ha desarrollado preferencias y capacidades específicas para manejar estos formatos. VOSviewer, por ejemplo, ha optimizado su motor para procesar archivos RIS manteniendo intactas las relaciones entre documentos, lo que es esencial para construir redes de co-citación precisas. CiteSpace, con su enfoque en análisis temporales, demanda estructuras de datos que preserven las cronologías de citación, algo que los formatos plain text derivados de Web of Science proporcionan naturalmente.

Bibliometrix en R representa quizás el punto más flexible en este espectro, capaz de ingerir múltiples formatos pero con una clara ventaja cuando se alimenta con RIS. Su función `convert2df()` no solo interpreta los metadatos estándar, sino que aplica algoritmos de limpieza y normalización que preparan los datos para análisis complejos como el mapeo de co-palabras o el cálculo de indicadores de colaboración internacional.

Para los investigadores que trabajan con pipelines personalizados, Python ofrece librerías como PyBibX que puentean las limitaciones de BibTeX para análisis bibliométricos. Estas herramientas permiten, por ejemplo, extraer redes semánticas desde conjuntos de referencias originalmente destinados a la composición de documentos académicos. No obstante, el proceso requiere frecuentemente etapas intermedias de transformación, especialmente cuando los datos provienen de múltiples fuentes con diferentes estándares de metadatos.

Las decisiones sobre formatos y conversiones repercuten directamente en la validez de los hallazgos bibliométricos. Un caso ilustrativo es el análisis de colaboraciones institucionales: mientras RIS preserva típicamente la estructura completa de afiliaciones, las conversiones automáticas a CSV pueden fragmentar esta información en columnas separadas, introduciendo artefactos en los mapas de redes generados por herramientas como CitNetExplorer.

La experiencia recomienda adoptar un enfoque estratificado para la gestión de formatos. En la fase inicial de recolección, RIS suele ser la opción más segura, garantizando que ningún metadato relevante se pierda durante la exportación desde las bases de datos. Para las etapas de limpieza y normalización, CSV brinda la flexibilidad necesaria para manipular grandes volúmenes de datos con herramientas estadísticas. Finalmente, en la fase de visualización y análisis especializado, muchos investigadores optan por regresar a formatos como RIS o a estructuras nativas del software bibliométrico utilizado.

Este ciclo iterativo subraya un principio fundamental: en bibliometría avanzada, los formatos de intercambio de datos no son meros contenedores pasivos, sino componentes activos que condicionan las posibilidades analíticas. La creciente interoperabilidad entre herramientas, impulsada por estándares como OpenAPI y formatos como JSON-LD, promete simplificar estos flujos de trabajo en el futuro. Sin embargo, hasta que estas innovaciones se adopten masivamente, el conocimiento profundo de RIS, BibTeX y CSV sigue siendo una habilidad esencial para cualquier investigador que busque extraer perspectivas significativas de los datos bibliográficos.

Ya está toda la información, pero ahora, ¿qué hacer con ella? En la próxima sección se abordará todo el tratamiento de los datos para obtener resultados e interpretaciones de alto nivel.

Recapitulando

- La obtención de datos bibliométricos es la primera fase práctica de un estudio y

define la validez de todo el análisis posterior.

- Las principales fuentes de información son: Scopus, Web of Science (WoS) y PubMed, por su cobertura, estructura y exportabilidad de metadatos.
- En Scopus, la búsqueda se optimiza usando operadores booleanos (AND, OR, NOT), filtros por año, país, área o afiliación, y códigos SOURCE-ID.
- En Web of Science, se emplean las colecciones *Core Collection* e *Emerging Sources* y las categorías temáticas para refinar resultados.
- En PubMed, se utilizan términos controlados MeSH, límites clínicos y estrategias de combinación entre título/resumen/palabras clave.
- Las herramientas alternativas amplían la obtención de datos cuando las bases tradicionales son insuficientes o restringidas. Entre ellas destacan:
    - Scholar (scraping ético) → útil para ampliar cobertura pero con cautela por ruido y duplicados.
    - Harzing's Publish or Perish → permite extraer y analizar métricas de Google Scholar con filtros de Google personalizables.
    - Gestores de referencias (Zotero, Mendeley) → facilitan la exportación masiva de metadatos y conversión de formatos.
- Los formatos más usados para exportar registros son:
    - RIS → compatible con la mayoría de softwares bibliográficos.
    - BibTeX → estándar para análisis en R o Python.
    - CSV → permite manipulación directa en Excel o software estadístico.
- Es fundamental mantener la coherencia del formato durante el proceso, evitando pérdida de campos (autores, afiliación, DOI, país, año).
- Se recomienda conservar la versión original de cada *dataset* y documentar el flujo de obtención de datos (fecha, fuente, filtros y número final de registros).
- La ética en la obtención de datos prohíbe prácticas como *scraping* masivo sin autorización o el uso de datos con derechos restringidos.
- Se promueve la reproducibilidad mediante la descripción clara de los criterios de búsqueda y la publicación del *dataset* o código empleado.
- El capítulo vincula la obtención de datos con las siguientes etapas metodológicas: limpieza, análisis y comunicación de resultados.
- En esta fase se definen los datos pertinentes: variables bibliográficas (autor, año, revista, país, citas, afiliación, tipo de documento).
- Los errores más comunes en esta etapa son: duplicados, registros incompletos, errores en nombres de autores y confusión entre fuentes.
- El control de calidad incluye verificar consistencia, eliminar duplicados y homogeneizar nombres mediante normalización.
- La documentación completa del proceso de búsqueda aumenta la transparencia y credibilidad del estudio bibliométrico.
- Se sugiere acompañar esta fase con capturas de pantalla o scripts reproducibles (ej. en Google Colab o RMarkdown).
- El capítulo enfatiza que sin una base de datos sólida, las métricas y visualizaciones posteriores pierden fiabilidad.
- En síntesis, esta parte enseña a construir el "corazón empírico" del estudio: una base bibliográfica ética, limpia, completa y normalizada.

Preguntas de autoevaluación

1. ¿Por qué la obtención de datos es considerada la fase crítica en un estudio bibliométrico?
2. ¿Qué características diferencian a Scopus, Web of Science y PubMed como fuentes

de datos?

3. ¿Qué función cumplen los operadores booleanos en las búsquedas bibliográficas?
4. ¿Qué ventajas y limitaciones presenta el uso de Google Scholar mediante *scraping* ético?
5. ¿Para qué sirve el programa *Harzing's Publish or Perish* en la bibliometría?
6. ¿Qué diferencia existe entre los formatos RIS, BibTeX y CSV al exportar registros?
7. ¿Qué buenas prácticas deben seguirse para asegurar la reproducibilidad de las búsquedas bibliográficas?
8. ¿Cuáles son los principales errores que pueden afectar la calidad del *dataset* bibliométrico?
9. ¿Qué criterios éticos deben respetarse en la recolección de datos científicos?
10. ¿Por qué la documentación y normalización del *dataset* son esenciales antes del análisis bibliométrico?

BIBLIOGRAFÍA

1. Moed HF. Citation analysis in research evaluation. Dordrecht: Springer; 2005. https://doi.org/10.1007/1-4020-3714-7

2. Thelwall M. Web indicators for research evaluation: a practical guide. San Rafael (CA): Morgan & Claypool Publishers; 2016. https://doi.org/10.2200/S00733ED1V01Y201602ICR048

3. Sugimoto CR, Larivière V. Measuring research: what everyone needs to know. Oxford: Oxford University Press; 2018. https://global.oup.com/academic/product/measuring-research-9780190640125

4. Ding Y, Rousseau R, Wolfram D, editors. Measuring scholarly impact: methods and practice. Cham (Switzerland): Springer; 2014. https://doi.org/10.1007/978-3-319-10377-8

5. Cronin B, Sugimoto CR, editors. Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact. Cambridge (MA): MIT Press; 2014. https://doi.org/10.7551/mitpress/9780262026792.001.0001

6. Cooper HM, Hedges LV. Research synthesis and meta-analysis: a step-by-step approach. 5th ed. Los Angeles (CA): SAGE Publications; 2019.

REFERENCIAS BIBLIOGRÁFICAS

1. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews. 2022; 18(2):e1230.