Chapter 4 / Capítulo 4

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.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

From Chaos to Organized Knowledge

4.1. Search Strategies for Bibliometric Studies

In bibliometrics, the quality of results depends fundamentally on a rigorous, well-planned search strategy. A poor approach at this initial stage can compromise the entire study, generating significant biases, omitting essential literature, or incorporating irrelevant material that distorts the analysis. The consequences of inadequate design are dire in this field, where the integrity of the findings depends on the comprehensiveness and accuracy of the documentary corpus retrieved.

This chapter offers a systematic methodological framework for developing optimal search strings on the leading bibliographic platforms (Web of Science, Scopus, PubMed, among others). The approach presented combines theoretical principles with practical applications.

- Advanced techniques for formulating balanced search equations.
- Validation strategies to ensure complete coverage without sacrificing accuracy.
- Concrete examples drawn from various scientific disciplines.
- Frequent errors and how to avoid them, based on analysis of real cases.

The proposed method emphasizes the need to adapt each search to the particularities of the documentary system of each field of study, considering factors such as specific disciplinary terminology, the indexing characteristics of each database, and the biases inherent in different information retrieval systems.

Through this structured approach, researchers will be able to build robust document sets that serve as a reliable basis for advanced bibliometric analysis, from scientific mapping studies to impact assessments. The methodology presented has been empirically validated across multiple research projects, demonstrating its effectiveness in producing representative, replicable results across various areas of knowledge.

4.1.1. Designing the search strategy

Defining the scope: the basis of an effective bibliometric search

The first critical step in any bibliometric study is to establish the research's scope and objectives precisely. This delimitation process requires answering fundamental questions that will determine the entire subsequent methodological design.

The central research question must be clearly formulated, as it will define the type of analysis to be performed. Is the objective to identify temporal trends in a field? To evaluate the impact of certain institutions? To map international collaboration networks? Each of these purposes will require different search strategies.

Inclusion criteria are essential filters for ensuring the relevance and manageability of the results:

- The time period must be justified according to developments in the field (e.g., in emerging areas, the last 5 years may be sufficient; in established disciplines, a decade or more may be necessary).
- The types of documents selected (articles, reviews, patents, theses) must be aligned with the communication practices of each discipline.
- The decision on languages involves a balance between comprehensiveness (including

contributions in multiple languages) and feasibility (the predominance of English in many scientific areas).

Applied example: To study "*Scientific production on artificial intelligence applied to medicine between 2015-2025,"* an appropriate delimitation could specify:

- Only full articles in peer-reviewed journals
- Documents indexed in major databases (WoS, Scopus)
- Publications in English (given its predominance in these technical disciplines)
- Exclusion of patents and gray literature to maintain focus on basic research

This planning phase, although often underestimated, is crucial. A large part of the errors in bibliometrics originates from an imprecise definition of the initial scope. A clear delimitation subsequently avoids the inclusion of documentary noise or the omission of key literature, problems that can compromise the validity of the findings.

Select databases

Given the planned scope, the database to be used for the search must be selected, which requires basic knowledge of each database's content and strengths. Each database has advantages and biases:

Table 4.1. Coverage, filters, and limitations of the main scientific databases

Database	Coverage	Useful filters	Limitations
Scopus	+25,000 journals, strong in STEM	Integrated citation analysis	Expensive, weak in humanities
Web of Science	+20,000 journals, historical core	Built-in h-index	Bias toward the US/Europe
PubMed	+30 million records in biomedicine	Filters by study type (clinical trials)	No complete citation data
Google Scholar	Free, includes books and gray literature	Semantic search	No analysis tools
Dimensions	Integrates publications and patents	Funding data	Uneven coverage by region

Database biases can be offset by using multiple databases (e.g., Scopus + PubMed).

4.1.2. Building search strings

Once the topic of the search has been clearly defined, the words most representative of the research should be selected.

Keywords and Boolean operators

The effectiveness of a bibliometric search depends on how key terms are combined using logical operators. The AND operator restricts results by connecting concepts that must appear simultaneously (example: *"machine learning"* AND *"healthcare"* retrieves only documents that mention both topics). Conversely, OR broadens the scope by including synonyms or conceptual variants (e.g., "*deep learning"* OR "*neural networks"*). NOT is used to exclude irrelevant topics, although it should be used with caution so as not to eliminate useful literature (e.g., "*cancer*" NOT *"breast cancer"* could omit critical cross-sectional studies).

Quotation marks ensure exact searches for complete phrases, avoiding scattered results

("climate change" is more accurate than climate AND change). Wildcards () capture morphological variants of a common root (e.g., "education," "educative," "educational").

Example of an optimized string

("artificial intelligence" OR "AI") AND ("medical diagnosis" OR "clinical decision") NOT "robotics."

This query retrieves documents on AI applied to medical diagnosis, excluding those focused on robotics.

The above operators are common to most databases and search engines, but some databases add new ones to increase search specificity. In any case, the logic behind these operators is the same regardless of where they are implemented.

Advanced filters for precision

Often, the above is not enough to retrieve the required or targeted content, or it covers much more than planned. Databases allow you to refine searches using specific filters:

- By field: in Scopus, TITLE-ABS-KEY() searches the title, abstract, and keywords; in PubMed, TI/AB covers the title and abstract. Limiting the search to the title only (TITLE()) underestimates the relevant literature, as many key concepts may only appear in the abstract.
- By impact: filters such as CITES > 50 (WoS) identify highly cited works, useful for studies of scientific influence.
- By institution: AFFIL("Harvard University") locates the output of a specific university, essential for collaboration analysis.

Standard error and solution

Problem: using only TITLE() can cause relevant documents to be missed

Solution: always combine title, abstract, and keywords (TITLE-ABS-KEY() in Scopus or TI/AB/KW in WoS).

4.1.3. Specific techniques by database

In Scopus, advanced syntax allows precise filtering using operators such as TITLE-ABS-KEY(), which covers titles, abstracts, and keywords simultaneously. For example, a query such as TITLE-ABS-KEY("blockchain" AND "supply chain") AND PUBYEAR > 2015 retrieves only recent publications on blockchain applications in supply chains.

Web of Science offers unique features for institutional network analysis. Its TS=() syntax for search terms combined with AD=() for affiliations allows you to map collaborations between specific research centers. A query such as TS=("quantum computing") AND AD=("MIT" OR "Stanford") identifies the joint output of these universities in quantum computing.

For biomedical research, PubMed stands out for its controlled vocabulary system, MeSH (Medical Subject Headings). This hierarchical thesaurus resolves problems of synonymy and terminological ambiguity. A search such as "COVID-19"[Mesh] AND "Vaccines"[Mesh] AND "clinical trials"[Publication Type] ensures complete coverage of COVID-19 vaccine clinical trials, avoiding the biases of free-text searches. PubMed's precision makes it indispensable for systematic reviews in health sciences.

Each platform requires specific strategies to optimize precision and recall. Scopus and WoS

prioritize control over terminology and integrated analytical capabilities, while PubMed offers standardized vocabulary, and Google Scholar provides documentary breadth at the expense of methodological transparency. The choice between them should be based on the objectives of the study: citation analysis and collaboration (WoS/Scopus), rigorous clinical reviews (PubMed), or identification of unconventional literature (Google Scholar). A robust research design often combines multiple databases to compensate for their respective limitations.

Recently, tools such as Dimensions have emerged as promising alternatives, integrating publications, patents, and datasets into a single environment with an open API. Their flexible syntax (e.g., search publications for "CRISPR" where year > 2020) and multidisciplinary coverage are redefining the standards for complex bibliometric searches. However, the widespread adoption of these new systems still faces interoperability challenges with traditional metrics.

In summary, the specific operators and filters for each database are:

Scopus

Boolean and advanced search operators

- AND, OR, AND NOT → Combine terms.
- → Truncation (e.g., comput searches for "computer," "computing," etc.).
- " " → Exact search (e.g., "machine learning").
- W/n → Words within n terms of distance (e.g., AI W/3 healthcare).
- PRE/n → First word before the second, with n terms away (e.g., digital PRE/2 transformation).

Advanced filters

- Year: PUBYEAR > 2020
- Document type: DOCTYPE(ar) (articles) or DOCTYPE(re) (reviews).
- Specific field:
  - TITLE-ABS-KEY("blockchain") → Search in title, abstract, or keywords.
  - AFFIL("Harvard University") → By affiliation.
- Citation indexes: REFERENCES() (e.g., REFERENCES(123456789)).

Combined example: TITLE-ABS-KEY("quantum computing") AND PUBYEAR > 2018 AND AFFIL("MIT") AND DOCTYPE(ar)

Web of Science (WoS)

Operators and syntax

- AND, OR, NOT → Standard operators.
- → Truncation (e.g., bio for "biology," "biotechnology").
- " " → Exact phrase.
- NEAR/x → Nearby terms (e.g., climate NEAR/2 change).
- SAME → Terms in the same field (useful for addresses: UNIV COLORADO SAME BOULDER).

Advanced filters

- WC categories: WC=("Computer Science").
- Year: YR=2020-2023.
- Document type: DT=("Article").
- Search by field:
  - TS=("neural network") → Subject (title, abstract, keywords).

- AU=(“Smith J”) → Author.
- OG=(“Stanford University”) → Organization.

Combined example: TS=(“CRISPR” AND “gene editing”) AND PY=(2020-2023) AND WC=(“Biotechnology”) AND DT=(“Review”)

PubMed

Operators and syntax

- AND, OR, NOT → Boolean operators.
- → Truncation (e.g., immun for “immune,” “immunity”).
- “ “ → Exact phrase.
- [Field] → Search by field (e.g., “COVID-19”[Title]).

Advanced filters (using tags)

- Article type: “review”[Publication Type].
- Date: “2020/01/01”[Date - Publication] : “2023”[Date - Publication].
- Specific fields:
  - “deep learning”[Title/Abstract].
  - “NIH”[Affiliation].
  - “1AU”[Author] → Lead author.
- Predefined filters:
  - “free full text”[Filter].
  - “clinical trial”[Filter].

Combined example: “artificial intelligence”[Title/Abstract] AND “diagnosis”[Title]) AND “2020”[Date - Publication] AND “systematic review”[Publication Type]

4.1.4. Validation and documentation: ensuring rigor in bibliometric searches

The validation phase is essential to ensure that the results of a bibliometric search are comprehensive, accurate, and reproducible. A rigorous process begins with a **pilot test**, where the first 50-100 records obtained are critically analyzed. This initial sampling allows frequent problems to be identified: overly broad terms that generate documentary noise, or, conversely, excessively restrictive search strings that omit relevant literature. Iteration is key here; each adjustment to the Boolean operators or term selection must be tested until an optimal balance between precision and sensitivity is achieved.

To ensure methodological transparency, the **PRISMA** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) **protocol** provides a standardized framework applicable to bibliometric studies. Its systematic approach requires detailed documentation of the entire selection process: the total number of records identified in each database, the exclusion criteria applied (duplicates, documents outside the temporal or thematic scope), and the decisions made at each stage of filtering. The PRISMA flowchart thus becomes a powerful visual tool that summarizes the process of refining the document corpus, allowing other researchers to evaluate the methodological soundness of the study.[1]

Detailed documentation should also include: the exact syntax of all search strings used (including discarded variants), the dates on which the queries were run (important given the dynamic nature of databases), and any known limitations in the coverage of the sources used. Clarity on all of these points is crucial from the initial phase of the research, as, as will be seen in subsequent chapters, all of this must be detailed in the final report.

Experienced researchers often supplement this process with **external validation**, consulting subject matter experts to verify that the search has not omitted seminal contributions, or comparing their results with existing systematic reviews in the field. This additional step helps to detect possible biases in the search strategy that could affect the study's conclusions. The combination of internal validation (using standardized protocols) and external validation (through expert judgment) is the best guarantee of producing robust and reliable bibliometric results.

This methodical approach transforms the bibliographic search from a mere technical procedure into a rigorous research process, where each decision is justified and documented. The time invested in this critical phase directly affects the quality and credibility of subsequent findings, helping avoid the biases and limitations that often affect bibliometric studies conducted without this systematic approach.

This progressive reduction must be meticulously documented, specifying the reasons for exclusion at each stage. Complete traceability of the process not only strengthens the internal validity of the study but also facilitates its replication or future updating.

4.1.5. Common errors and solutions

An overly broad search, yielding more than 10,000 records, is often unmanageable for detailed bibliometric analysis and may include substantial documentary noise. This problem frequently occurs when using generic terms without adequate filters. The solution is to incorporate strategic restrictions such as time limits (e.g., last 10 years), filtering by document type (research articles only), or focusing on specific fields (title and abstract rather than full text).

At the other extreme, an overly narrow search can miss seminal contributions by using overly specific terminology. This is particularly problematic in emerging fields where concepts have not yet standardized their nomenclature. To avoid this, it is essential to build search strings that incorporate synonyms and terminological variants using OR operators, in addition to consulting disciplinary thesauri such as MeSH in PubMed or Scopus controlled terms.

A common mistake in specific disciplines is to ignore books and conference proceedings, which creates a systematic bias against areas such as the humanities and computer science. Traditional databases (WoS/Scopus) have limited coverage in these formats, so the search should be supplemented with Google Scholar for books or Dimensions for proceedings. In interdisciplinary studies, this omission can lead to missing much of the relevant literature.

Failure to document the search strategy compromises the study's reproducibility, a fundamental principle of research. The solution involves systematically saving not only the final chain but all its iterations, specifying the platform used, the search date, the filters applied, and the number of results obtained at each step. Standard formats such as PRISMA diagrams or systematic review protocol templates can be adapted for this purpose. Tools such as Zotero or Mendeley allow search records to be exported with complete metadata for later auditing.

These problems and their solutions illustrate how the design of bibliometric searches is an iterative process that requires a balance between comprehensiveness and precision. Experience shows that dedicating 30 % of the total study time to refining and validating the search strategy can significantly improve the quality of the final results. Incorporating expert peer reviews in the methodological design phase helps identify potential omissions or biases before executing

the final searches.

The growing use of machine learning techniques for automated screening of large volumes of results is transforming this process, but the initial construction of accurate search strings remains essential. Platforms such as ASReview or Rayyan can then assist in the selection phase, but they depend critically on a well-designed document retrieval strategy from the outset. This synergy between traditional methods and new technologies paves the way for more efficient and comprehensive bibliometric searches.

4.2. Alternative tools for bibliometric search: broadening horizons

Although Scopus and Web of Science remain benchmarks in bibliometric analysis, there are valuable alternatives that democratize access to scientific information. Google Scholar, with its extensive coverage of gray literature, along with tools such as Publish or Perish and advanced reference managers, offers unique possibilities for researchers with limited resources or who work with unconventional formats.

4.2.1. Google Scholar: practical advantages and limitations

The main strength of Google Scholar lies in its ability to index content that is often left out of commercial databases: from preprints on arXiv to doctoral theses and working papers. This comprehensiveness makes it particularly useful for studies in the humanities or emerging disciplines where formal publication can be slower. However, its systematic use presents significant technical challenges. The approximate limit of 100 queries per day before IP blocking forces the use of careful sampling strategies, while the lack of structured metadata complicates advanced bibliometric analysis.

For researchers with technical skills, Python's scholarly library offers a balance between automation and compliance with terms of service. The basic process involves: installation via pip, configuration of specific queries, and controlled data extraction. A script might include:

```python
import csv
from scholarly import scholarly
search_query = scholarly.search_pubs("blockchain in healthcare")
with open('results.csv', 'w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerow(["Title", "Authors", "Year", "Citations"])
    for i in range(50):   Responsible limit
        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
```

This approach includes key safeguards: record limits, error handling, and CSV storage for later

analysis. For non-programmers, extensions such as Scholarcy offer basic extraction functionality directly from the browser, albeit with less customization.

Academic data scraping must follow clear principles: intervals between queries that simulate human behavior, respect for each site's robots.txt, and exclusive use for research purposes. It is crucial to remember that many documents on Google Scholar are protected by copyright, so mass extraction of full texts without permission constitutes an ethical violation. Best practices recommend:

- Limiting extractions to basic metadata (title, authors, citations).
- Storing only the information necessary for analysis.
- Properly citing all retrieved sources.
- Periodically checking the platforms' terms of service.

These alternative tools work best as complements, not replacements, to traditional databases. A robust workflow could begin with systematic searches in Scopus/WoS, followed by a sweep of Google Scholar to capture additional literature, and end with tools such as Zotero or Mendeley for organization and deduplication. This methodological triangulation mitigates the inherent biases of each platform while maximizing document coverage.

4.2.2. Harzing's Publish or Perish

Publish or Perish (PoP) not only extracts data from Google Scholar but also integrates with key academic databases such as Scopus, Web of Science (WoS), and PubMed, significantly expanding its usefulness for researchers who require more structured analysis. This multi-platform capability makes it a versatile solution for comparative bibliometric studies, allowing impact metrics to be compared across different sources and detecting possible discrepancies in citation coverage. Chapter 3 discussed in detail how to install this tool and its advantages and features. In general, to perform a search with it, follow these steps:

1. Download and installation:
    - Available at https://harzing.com/resources/publish-or-perish (Windows/Mac/Linux).

2. First search:
    - Open the software and click on "New Query."
    - Enter search terms (e.g., "artificial intelligence education").

3. Export results:
    - Go to File > Save As and choose CSV/Excel for analysis in Bibliometrix or Excel.

4.2.3. Reference managers for advanced bibliometric analysis

Reference managers such as Zotero and Mendeley have evolved beyond simple bibliographic organizers into powerful tools for extracting and mass-processing bibliometric data. Their ability to capture structured metadata directly from academic portals or PDF files makes them indispensable in quantitative research workflows.

Zotero stands out for its seamless integration with databases such as Scopus and Web of Science. Using its browser connector, researchers can import hundreds of references with a single click, preserving critical information such as DOIs, institutional affiliations, and citation fields. The CSV export process (via the context menu) generates tables ready for analysis in specialized software. However, care is required when selecting custom fields to ensure that all

relevant metadata is included.

Mendeley offers distinct advantages, particularly in handling existing PDF collections. Its document recognition engine can extract metadata even from files without embedded information, although with variable accuracy.

The "Verify Document Details" feature allows you to correct errors commonly found in older articles or less established journals. For bibliometric projects, its ability to export in RIS format is especially valuable, as this standard preserves relationships between documents that tabular formats such as CSV might lose.

A recurring challenge with both tools is the inconsistent quality of metadata. Recent studies show that up to 15 % of articles in reference managers require manual correction of basic fields, such as the year of publication or the complete list of authors. Here, APIs such as CrossRef become essential add-ons.

The Python example shown (using the habanero library) illustrates how to automate the retrieval of missing metadata via queries based on titles or text fragments. This hybrid approach—combining automatic extraction with programmatic verification—can significantly improve the efficiency of preparing bibliometric data.

Advanced technical considerations arise when working with large volumes of references. Zotero allows custom scripts to be run through its internal API (Zotero API), enabling tasks such as:

- Automatic normalization of institutional names.
- Detection and merging of duplicates based on semantic similarity.
- Extraction of citations between documents through reference analysis.

Meanwhile, Mendeley Desktop (unlike its web version) provides direct access to its local SQLite database, enabling complex queries beyond the limitations of the graphical interface. Researchers with SQL knowledge can thus perform advanced data transformations before final export.

The choice between these tools often depends on the specific workflow. Zotero shines in collaborative projects thanks to its cloud synchronization and granular permissions, while Mendeley may be preferred in environments where managing large volumes of PDFs is a priority. Both, however, share the potential to serve as a bridge between initial literature collection and sophisticated bibliometric analysis on platforms such as VOSviewer or Bibliometrix.
These managers can be used to extract data from PDFs or even from the web for later use in bibliometric studies.

4.3. Export formats in bibliometrics: interoperability and analysis

Selecting the export format in bibliometrics is a critical decision that directly affects the quality and depth of subsequent analysis. The RIS (Research Information Systems) format, initially developed for academic databases, stands out for its ability to preserve complex metadata such as citation relationships and hierarchical authorship structures. This format maintains a unique balance between human readability and automated processing, making it particularly valuable when working with network analysis tools such as VOSviewer. Its standardized tag structure (e.g., AU for authors or PY for year of publication) enables a smooth transition between platforms without loss of critical information.

BibTeX, on the other hand, offers distinct advantages in environments where technical academic writing is paramount. Its minimalist syntax and native integration with LaTeX systems make it irreplaceable for researchers who need to manage references while writing. However, this same simplicity becomes a limitation when addressing advanced bibliometric analysis, where richer metadata on citations, institutional affiliations, or impact indicators is required. Modern tools such as PyBibX have expanded their usefulness by enabling some bibliometric processing, but they still need additional transformations to achieve the level of detail provided by other formats.

The CSV format represents the most direct bridge to advanced quantitative analysis. Its tabular structure is ideally suited to statistical environments such as R (via Bibliometrix) or Python (with Pandas), where manipulating large volumes of data requires flexibility. However, this same adaptability carries risks: automatic exports from some platforms can flatten complex data structures, such as author lists or institution hierarchies, into text strings that require further processing. The key to leveraging CSV in bibliometrics is to ensure that the initial export captures all relevant metadata in well-defined columns.

4.3.1. Conversion between formats: tools and uses in bibliometric software

The reality of contemporary bibliometric work often requires navigating between multiple formats, each optimized for different stages of the research flow. Conversions between RIS, BibTeX, and CSV are not mere technical translations, but processes that can expand or limit analytical possibilities. Tools such as Zotero perform these transformations acceptably for basic uses, but large-scale projects require more sophisticated approaches.

In today's bibliometric software ecosystem, each primary tool has developed specific preferences and capabilities for handling these formats. VOSviewer, for example, has optimized its engine to process RIS files while preserving document relationships, which is essential for building accurate co-citation networks. CiteSpace, with its focus on temporal analysis, demands data structures that preserve citation chronologies, which plain-text formats derived from the Web of Science naturally provide.

Bibliometrix in R sits at the most flexible end of this spectrum, capable of ingesting multiple formats but with a clear advantage when fed RIS. Its `convert2df()` function not only interprets standard metadata, but also applies cleaning and normalization algorithms that prepare the data for complex analyses such as co-word mapping or the calculation of international collaboration indicators.

For researchers working with custom pipelines, Python offers libraries such as PyBibX that bridge the limitations of BibTeX for bibliometric analysis. These tools, for example, allow the extraction of semantic networks from sets of references intended initially for the composition of academic documents. However, the process often requires intermediate transformations, especially when the data comes from multiple sources with different metadata standards.

Decisions about formats and conversions directly impact the validity of bibliometric findings. An illustrative case is the analysis of institutional collaborations. While RIS typically preserves the complete structure of affiliations, automatic CSV conversions can fragment this information into separate columns, introducing artifacts in network maps generated by tools such as CitNetExplorer.

Experience recommends adopting a layered approach to format management: in the initial

collection phase, RIS is usually the safest option, ensuring that no relevant metadata is lost during database exports. For the cleaning and normalization stages, CSV provides the flexibility needed to manipulate large volumes of data with statistical tools. Finally, in the visualization and specialized analysis phase, many researchers choose to return to formats such as RIS or to structures native to the bibliometric software used.

This iterative cycle underscores a fundamental principle: in advanced bibliometrics, data exchange formats are not mere passive containers, but active components that condition analytical possibilities. The growing interoperability between tools, driven by standards such as OpenAPI and formats such as JSON-LD, promises to simplify these workflows in the future. However, until these innovations are widely adopted, a thorough knowledge of RIS, BibTeX, and CSV remains an essential skill for any researcher seeking to extract meaningful insights from bibliographic data.

All the information is already there, but now what to do with it? The following section will address the entire data processing workflow to obtain high-level results and interpretations.

Recap

- Obtaining bibliometric data is the first practical phase of a study and defines the validity of all subsequent analysis.
- The primary sources of information are Scopus, Web of Science (WoS), and PubMed, due to their coverage, structure, and metadata exportability.
- In Scopus, the search is optimized using Boolean operators (AND, OR, NOT), and results are filtered by year, country, area, or affiliation, as well as by SOURCE-ID codes.
- In Web of Science, the Core Collection, Emerging Sources, and subject categories are used to refine results.
- In PubMed, controlled MeSH terms, clinical limits, and combination strategies between title/abstract/keywords are used.
- Alternative tools expand data collection when traditional databases are insufficient or restricted. These include:
    - Scholar (ethical scraping) → helpful in expanding coverage, but caution is advised due to noise and duplicates.
    - Harzing's Publish or Perish → allows you to extract and analyze Google Scholar metrics with customizable Google filters.
    - Reference managers (Zotero, Mendeley) → facilitate bulk export of metadata and format conversion.
- The most commonly used formats for exporting records are:
    - RIS → compatible with most bibliographic software.
    - BibTeX → standard for analysis in R or Python.
    - CSV → allows direct manipulation in Excel or statistical software.
- It is essential to maintain format consistency throughout the process, avoiding loss of fields (authors, affiliation, DOI, country, year).
- It is recommended to retain the original version of each dataset and document the data acquisition process (date, source, filters, and the final number of records).
- Ethical data collection prohibits practices such as unauthorized mass scraping or the use of restricted data.
- Reproducibility is promoted by clearly describing the search criteria and publishing the dataset or code used.
- The chapter links data collection with the following methodological stages: cleaning, analysis, and communication of results.

- In this phase, the relevant data are defined: bibliographic variables (author, year, journal, country, citations, affiliation, document type).
- The most common errors at this stage are: duplicates, incomplete records, errors in author names, and confusion between sources.
- Quality control includes verifying consistency, eliminating duplicates, and standardizing names through normalization.
- Complete documentation of the search process increases the transparency and credibility of the bibliometric study.
- It is recommended to accompany this phase with screenshots or reproducible scripts (e.g., in Google Colab or RMarkdown).
- The chapter emphasizes that without a solid database, subsequent metrics and visualizations lose reliability.
- In summary, this part teaches how to build the "empirical heart" of the study: an ethical, clean, complete, and standardized bibliographic database.

Self-assessment questions

1. Why is data collection considered the critical phase in a bibliometric study?
2. What characteristics differentiate Scopus, Web of Science, and PubMed as data sources?
3. What role do Boolean operators play in bibliographic searches?
4. What are the advantages and limitations of using Google Scholar through ethical scraping?
5. What is the purpose of Harzing's Publish or Perish program in bibliometrics?
6. What is the difference between RIS, BibTeX, and CSV formats when exporting records?
7. What best practices should be followed to ensure the reproducibility of bibliographic searches?
8. What are the main errors that can affect the quality of the bibliometric dataset?
9. What ethical criteria should be respected when collecting scientific data?
10. Why are documentation and standardization of the dataset essential before bibliometric analysis?

BIBLIOGRAPHY

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

2. Thelwall M. Web indicators for research evaluation: A practical guide. San Rafael (CA): Morgan & Claypool Publishers; 2016. doi: 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. doi: 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. doi: 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.

BIBLIOGRAPHIC REFERENCES

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.