Chapter 2 / Capítulo 2

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch02

Brief Historical Overview / Breve Recuento Histórico

In 1955, a young chemist named Eugene Garfield published an article entitled *"Citation Indexes for Science,"* in which he proposed a system for tracking the impact of research. No one imagined that this idea would revolutionize science, giving rise to modern bibliometrics.(1)

2.1. The Analog Era (1960-1990): The Foundations

Science Citation Index (SCI) - 1960

In 1960, the Institute for Scientific Information (ISI), founded by Eugene Garfield, revolutionized scientific evaluation with the launch of the Science Citation Index (SCI), the first large-scale commercial citation index. Originally published in print, this system made it possible, for the first time, to systematically track how scientific articles were linked through their bibliographic references.(2)

The SCI was based on a simple but powerful principle: *"citations are connections of knowledge."* Its methodology included tracking cross-references among 600 selected scientific journals (mainly from the US and Europe), reverse indexing (instead of just listing authors or topics, the SCI allowed users to search for articles that cited a particular work, revealing its subsequent influence), and multidisciplinary coverage (although focused on the natural sciences, it laid the foundation for future indexes in the social sciences and humanities).

The SCI transformed the way scientific impact was measured. Before the SCI, productivity was measured by the number of publications. The index introduced the idea that an article could be influential even if its author published little (e.g., Watson and Crick's paper on DNA had few previous publications but changed biology). It identified highly cited works that defined entire fields (e.g., Miller's 1953 article on the origin of life). Despite its limitations, such as Anglophone bias, Western focus, and restricted access to privileged institutions, its influence endures through contemporary platforms such as Web of Science. At the same time, Google Scholar and Scopus have adopted their fundamental cross-citation logic.(3)

The SCI didn't just measure science: it made it more transparent and connected. Its history reminds us that even the most disruptive tools must be used with critical awareness.

Fun fact:

> *The original SCI occupied five linear meters of shelving, and only elite institutions such as Harvard could afford it.*(4)

Fundamental Laws (1970s-1980s). The mathematical pillars of bibliometrics

During the 1970s and 1980s, bibliometrics consolidated its scientific rigor through the validation and widespread application of two empirical laws formulated decades earlier: Lotka's Law (1926) and Bradford's Law (1934). These laws, although initially developed to describe patterns in scientific literature, became essential tools for understanding researcher productivity and the distribution of knowledge in academic journals.

Lotka's Law: inequality in scientific productivity

Lotka's Law, formulated by Alfred J. Lotka in 1926 but popularized during the scientometric boom of the 1970s, is one of the fundamental findings on the unequal distribution of scientific productivity. This principle states that approximately 10 % of researchers produce 50 % of academic publications, revealing a pattern of concentration of scientific output that transcends disciplines and historical periods. The law represents the specific application of the Pareto principle to the scientific domain, confirming that knowledge production follows an asymmetric distribution in which a few actors generate most of the research results.[5]

Contemporary empirical evidence consistently validates this unequal distribution across multiple fields of knowledge. This distribution enhances the "Matthew effect" in science, where established researchers with access to collaborative networks, institutional resources, and accumulated symbolic capital tend to publish more frequently, thus reinforcing their initial competitive advantages.

The practical applications of Lotka's Law in academic evaluation are numerous and significant. It enables the identification of highly productive researchers for hiring, promotion, or funding-allocation decisions, providing a quantitative reference point for evaluating exceptional performance. Simultaneously, it serves as a diagnostic tool for identifying structural inequalities within specific scientific systems, informing policies to democratize opportunities for publication and academic visibility.

However, the law has significant methodological limitations in contemporary contexts. Its applicability is more robust in STEM disciplines than in the humanities, where single authorship and diverse publication formats that escape traditional metrics predominate. Nor did it anticipate the explosion of massive co-authorship characteristic of large-team science, where publications in genomics or experimental physics can include thousands of authors, redistributing the dynamics of individual productivity. These limitations underscore the need to contextualize the law within the paradigmatic changes in scientific authorship and collaboration practices of the 21st century.

Bradford's Law: the core of key journals

Bradford's Law (1943) is a fundamental principle of bibliometrics that describes the uneven distribution of relevant scientific literature across academic journals. It establishes that a small core of periodicals concentrates the majority of significant articles in any field of knowledge, while a more extensive periphery hosts scattered contributions. This pattern of concentration reflects the specialized structure of communication in contemporary science, where specific journals serve as privileged channels for disseminating high-impact research.[6] Empirical evidence consistently confirms this phenomenon of concentration across multiple disciplines.

The practical applications of this law are particularly valuable for information resource management. Academic libraries use Bradford zone analysis to optimize subscriptions, focusing limited resources on the core of journals that maximize access to relevant literature. Simultaneously, the identification of "desert zones," fields where knowledge is widely dispersed among numerous sources, alerts us to the need for more comprehensive search strategies, especially in interdisciplinary areas or specific cultural studies.

However, Bradford's Law faces significant limitations in the contemporary scientific ecosystem. It has a marked Anglophone bias, with most articles published in large databases being in English, marginalizing valuable contributions in other languages. In addition, the emergence of open science and preprint repositories is substantially transforming these patterns of concentration. In artificial intelligence, for example, arXiv has disrupted traditional publishing by creating an alternative channel of dissemination that competes with established journals, demonstrating the growing fluidity of scientific communication patterns.

This evolution toward more distributed models suggests that, although the principle of concentration remains relevant, its concrete manifestations are rapidly changing. Contemporary bibliometrics must develop tools capable of capturing these new dynamics while maintaining the analytical utility of Bradford's original concept for understanding the structure of scientific communication.

General limitations of these laws

Although revolutionary, both laws have problems in the digital age:

1. Biased coverage: they were based on data from the US and Europe, ignoring scientific output from Asia, Africa, and Latin America. Example: Bradford's Law does not predict well the core of journals in African agronomy, where research is published in local journals.
2. Manual processing: in the 1970s and 1980s, counts were done by hand, which led to errors (e.g., duplication of authors with similar names).
3. Current context: today, algorithms such as those used by Scopus or Dimensions allow for more dynamic analysis, but the laws remain the theoretical basis.

Why do they still matter in the 21st century?

These historical bibliometric laws remain surprisingly relevant in the 21st-century digital scientific ecosystem, demonstrating that the fundamental patterns of academic communication transcend technological change. Lotka's Law continues to inform the development of modern evaluation systems, where the h-index and its variants incorporate its understanding of the uneven distribution of scientific productivity. This perspective contextualizes individual metrics within broader systemic patterns, preventing simplistic interpretations of research performance and recognizing the inherently asymmetrical nature of knowledge production.

In the realm of scientific policy, Bradford's Law provides a crucial analytical framework for navigating the exponential expansion of academic communication. Its principle of concentration helps identify predatory journals operating outside the legitimate cores of each discipline, offering an objective criterion for distinguishing reliable communication channels. This application is particularly valuable in open access contexts, where the proliferation of questionable publishers requires robust mechanisms to ensure the integrity of scientific communication.

The most profound influence of these laws is manifested in the algorithmic foundations that underpin contemporary digital tools. The PageRank algorithm of Google Scholar, for example, computationally implements Bradford's principle by assigning greater weight to citations from sources considered "nuclei" of academic authority. Simultaneously, scientific literature recommendation and discovery systems incorporate insights derived from Lotka to prioritize content from highly productive and influential researchers.

The continued validity of these principles demonstrates that, although scientific communication technologies have undergone radical transformations, the underlying patterns of knowledge production and dissemination maintain observable structural regularities. This continuity makes classical bibliometric laws essential tools for understanding both the continuities and transformations in contemporary scientific dynamics, providing an interpretive framework for navigating the complexity of today's research ecosystem.

2.2. The Digital Revolution (1990-2010): expansion and criticism

The Era in Which Bibliometrics Became Global (and Controversial)

The 1990s marked the beginning of **digital bibliometrics**, radically transforming how science is measured and managed. With the advent of the Internet, citation indexes moved away from print formats and adopted online platforms, expanding their reach but also generating new ethical and methodological challenges.

Web of Science (WoS) – 1997: The SCI goes digital

It was the digital version of *the Science Citation Index (SCI)*, launched by the **Institute for Scientific Information (ISI)**. For the first time, it enabled **real-time searches** and advanced citation analysis. It grew from 600 journals (in print) to 8,000 indexed journals, including social sciences (Social Sciences Citation Index) and arts (Arts & Humanities Citation Index). It introduced features such as *Citation Reports* (to calculate the impact of authors) and *the Journal Impact Factor (JIF) online* (previously available only in the printed *Journal Citation Reports*).[7]

The JIF, created in 1975 to evaluate journals, began to be misapplied to individual researchers. Universities require publication in “Q1 journals” for hiring, ignoring the actual quality of the articles.

Scopus (2004): the competitor that challenged the hegemony

Developed by Elsevier as a direct alternative to Web of Science (WoS), Scopus broadened the horizon of scientific indexing by incorporating a greater number of non-English-language journals, with a strong focus on European and Asian publications. Among its most significant contributions, it introduced author profiles and pioneered systems for name disambiguation that anticipated tools such as ORCID. It also incorporated alternative metrics such as CiteScore, conceived as a complementary indicator to the Journal Impact Factor (JIF).

However, Scopus was not without its critics. It has been accused of commercial bias, favoring Elsevier-associated publications in its rankings. Furthermore, its refusal to index preprints put it at a disadvantage compared to more open systems such as Google Scholar, as it excluded a significant portion of informal or “gray” scientific literature from its database.

Google Scholar (2004): democratic disruption

The launch of Google Scholar represented a profound transformation in access to scientific knowledge. Its automatic indexing system allowed it to incorporate not only peer-reviewed articles, but also preprints, such as those from arXiv, books, theses, and technical documents, many of which had remained off the radar of traditional databases. Unlike WoS and Scopus, Google Scholar was free, democratizing access to global scientific literature. Its ability to capture content in traditionally marginalized languages and formats, such as studies in Spanish from Latin America, gave it unprecedented coverage. In addition, with the development of Google Scholar Citations, researchers could create public profiles to showcase their academic output and personal metrics.

However, this openness also generated controversy. One of the main questions has been the opacity of its algorithms, as there is no information on how results are ranked, leading to strategies to manipulate the visibility of specific works through the intentional use of keywords. On the other hand, the quality of the indexed documents is a matter of debate, as Google Scholar includes predatory journals and non-peer-reviewed materials.

Impact and criticism of this revolution

The emergence of platforms such as Scopus and Google Scholar marked a turning point in the way scientific knowledge is accessed, evaluated, and disseminated. Among their main legacies is the globalization of science, by giving visibility to research from historically marginalized regions, such as Latin America, Asia, and Africa, beyond the traditional centers of academic production in the United States and Europe. In addition, the massive digitization of citations has enabled the development of new metrics, with the h-index, proposed by Hirsch in 2005, among the most influential for simultaneously measuring a researcher's productivity and impact.

However, structural problems persist that these platforms have not been able to resolve. Despite accumulated criticism, the impact factor (JIF) continues to be used as a decisive criterion in evaluation and funding processes, as is the case with national agencies such as ANID in Chile. [8] This persistence reinforces a quantitative logic that often ignores the quality or context of the knowledge produced. On the other hand, the digital divide remains a significant obstacle: many institutions in low- and middle-income countries cannot afford access to commercial databases such as WoS or Scopus, forcing them to rely on free alternatives such as Google Scholar, with all the limitations and risks that this implies in terms of reliability. Finally, the rise of digital tools has given rise to new forms of metric manipulation, such as citation stacking, a practice that consists of creating circles of fraudulent citations to artificially inflate the impact of specific publications, thus distorting academic evaluation processes.

2.3. The Modern Era (2010-Present): Complexity and Democratization

Altmetrics (2010)

Starting in 2010, bibliometrics began to expand into broader forms of scientific impact assessment, giving rise to altmetrics. This approach seeks to measure impact beyond the counting of academic citations, incorporating indicators such as mentions in social media, the media, blogs, Wikipedia, and public policy documents. The launch of Altmetric.com in 2011 marked a turning point, offering tools to track the digital circulation of scientific articles in real time.

Open Source Bibliometrics (2017)

In 2017, the democratization of bibliometrics took a significant leap forward with the emergence of open-source tools, which facilitated access to advanced analysis without the high costs of commercial platforms. Among the most notable are Bibliometrix, a R package for sophisticated bibliometric analysis, and VOSviewer, a widely used tool for visualizing co-authorship, co-citation, and keyword networks. These platforms enabled researchers at institutions with limited resources to access robust analytical methodologies, contributing to the decentralization of bibliometric knowledge.

Artificial Intelligence and Text Mining (2020)

In the 2020s, the integration of artificial intelligence revolutionized the way scientific literature is processed and analyzed. Models such as GPT-4 enable automated, efficient summarization of trends from millions of abstracts, allowing synthetic analysis of entire fields of research in record time. At the same time, platforms such as Dimensions.ai began to integrate databases that connect scientific publications with patents, grants, and research results, offering a much more comprehensive perspective on the cycle of knowledge production and transfer. This convergence between AI, text mining, and open science has expanded the scope and depth of contemporary bibliometrics, positioning it as a key tool in scientific and public policy decision-making.

Current challenge:

> *How to prevent AI from generating "zombie articles" (well-cited texts with no real substance).*

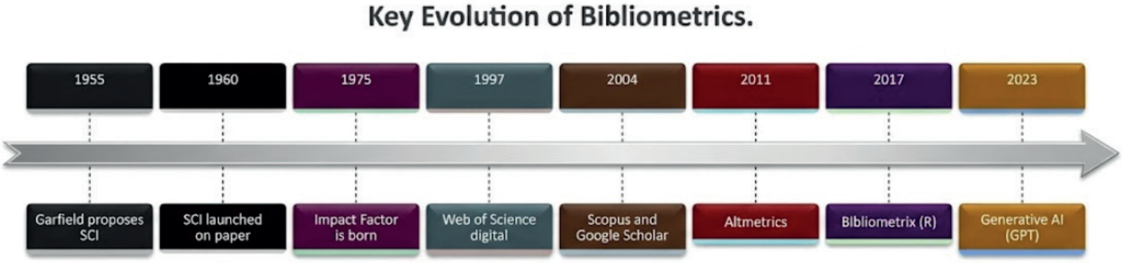

Figure 2.1. Visual timeline

2.4. Historical lessons

1. From analog to digital: Technological leaps have democratized access but introduced new biases (e.g., information overload).
2. From quantitative to qualitative: Criticism of IF led to responsible metrics (DORA, Leiden).
3. From academic to social: Altmetrics broadened the notion of scientific impact.

“Bibliometrics is no longer just about counting citations, but understanding how knowledge flows—and sometimes stagnates—in society.”

Section II will explore the most commonly used tools for constructing bibliometric analyses and how to obtain data for bibliometric studies practically.

Recap

- Bibliometrics has its origins in the first half of the 20th century, when the first attempts to quantify scientific production emerged.
- The term “bibliometrics” was coined by Paul Otlet (1934) and consolidated by Alan Pritchard (1969).
- Advances influenced its development in documentation, statistics, and information technology.
- The first studies on author productivity and article distribution were published in the 1920s and 1940s.
- Lotka (1926) formulated the Law of Scientific Productivity, which describes the unequal frequency with which researchers publish.
- Bradford (1934) proposed his Law of Dispersion, which explains how relevant articles are concentrated in a small number of core journals.
- Zipf (1949) introduced the Law of Word Frequency, which serves as the basis for the analysis of terms and co-occurrences.
- In the 1950s and 1960s, Eugene Garfield founded the Institute for Scientific Information (ISI) and developed the Science Citation Index (SCI).
- The SCI revolutionized the measurement of science by enabling the tracking of citation networks between publications.
- In the 1970s and 1980s, the field became institutionalized with the emergence of the journal Scientometrics and the first international conferences.
- At this stage, bibliometrics expanded into scientometrics, focusing on the dynamics and politics of science.
- Starting in the 1990s, the use of electronic databases and specialized software allowed for broader and more accurate analysis.

- The emergence of the Internet and search engines radically transformed access to and the collection of scientific information.
- With the arrival of Google Scholar (2004) and Scopus (2004), bibliometric sources and impact indicators diversified.
- Since 2010, altmetrics and webmetrics have emerged, focusing on the social and digital impact of science.
- In this modern stage, bibliometrics is integrated with artificial intelligence, big data, and open science.
- Current methods enable mapping collaboration networks, thematic trends, and cognitive structures.
- The historical development reflects a transition from a simple quantitative approach to a multidimensional and ethical view of scientific evaluation.
- Contemporary bibliometrics combines the tradition of Garfield, Lotka, and Bradford with advanced digital tools.
- In short, its evolution has made bibliometrics an essential pillar of research, knowledge management, and science policy.

Self-assessment questions

1. Who coined the term “bibliometrics” and in what year did its use become established?
2. What does Lotka’s Law describe in relation to scientific productivity?
3. What is the central principle of Bradford’s Law?
4. What did Zipf contribute to the development of bibliometrics?
5. How important was Eugene Garfield in the history of the discipline?
6. What role did the creation of the Science Citation Index play in the 1960s?
7. Why did the 1970s and 1980s mark the institutionalization of bibliometrics?
8. How did electronic databases and the Internet influence the evolution of the field?
9. What contributions did altmetrics introduce after 2010?
10. How does contemporary bibliometrics differ from its early stages?

BIBLIOGRAPHY

1.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

2. Glänzel W. Bibliometrics as a research field: A course on theory and application of bibliometric indicators. Leuven: KU Leuven; 2003. https://www.kuleuven.be/metaforum/docs/pdf/lecture_glanzel_bibliometrics.pdf

3. Waltman L. A review of the literature on citation impact indicators. J Informetrics. 2016;10(2):365–91. doi:10.1016/j.joi.2016.02.007

4. 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

5. Hood WW, Wilson CS. The literature of bibliometrics, scientometrics, and informetrics. Scientometrics. 2001;52(2):291–314. doi:10.1023/A:1017919924342

BIBLIOGRAPHIC REFERENCES

1. Garfield E. Citation indexes for science. Science. 1955;122(3159):108–11.

2. Garfield E. “Science Citation Index”—A new dimension in indexing. Science. 1964;144(3619):649–54.

3. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: a multidisciplinary comparison of coverage via citations. Scientometrics. 2021;126(1):871–906. https://doi.org/10.1007/s11192-020-03690-4

4. Garfield E. The evolution of the science citation index. International Microbiology. 2007;10(1):65–9.

5. Lotka AJ. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 1926;16(12):317–23.

6. Bradford SC. Classic paper: Sources of information on specific subjects. Collection Management. 1976;1(3–4):95–103. https://doi.org/10.1300/J105v01n03_06

7. Patsopoulos NA, Analatos AA, Ioannidis JPA. Relative citation impact of various study designs in the health sciences. JAMA. 2005;293(19):2362–6. https://doi.org/10.1001/jama.293.19.2362

8. Agencia Nacional de Investigación y Desarrollo (ANID). Concurso de Proyectos Fondecyt Regular 2024. 2024. https://anid.cl/concursos/concurso-de-proyectos-fondecyt-regular-2024/