Chapter 1

Introduction to Bibliometrics and Current Data Sources

What Is Bibliometrics?

A Very Short History

The term *bibliometrics* is widely attributed to Alan Pritchard from his 1969 paper titled "Statistical Bibliography or Bibliometrics" (Andrés 2009; Gingras 2016; Pritchard 1969). However, before the term was coined, bibliometrics was already emerging as a viable scientific discipline in the 1960s, in large part due to the foundation of the Institute for Scientific Information (ISI) led by Eugene Garfield and the subsequent development of the Science Citation Index (SCI; Mokhnacheva and Tsvetkova 2020). The intention in creating the SCI was to "eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making [scholars] aware of criticisms of earlier papers" (Garfield 1955). Later, the ISI recognized the power of the data available in the SCI for creating networks among journals and their citations and developed what is now the widely used (and disputed) Journal Impact Factor (JIF), the average citations per publication. The JIF rose in popularity at Garfield's suggestion that it would be helpful to librarians for managing library collections. However, less discussed in the literature are Garfield's other suggested applications, which include use by individual researchers for selecting reading lists, by editors for evaluating journal performance, and in the study of science policy and research evaluation (Garfield 1972). Although there have been mounting critiques on the limitations of the JIF, many of these described applications remain core to bibliometrics more broadly, even though the JIF may not be the metric of choice.

Defining Bibliometrics

Despite the continually evolving methods of analysis, the heart of bibliometrics remains the counting of documents, their related bibliographic information, and their network of citations. The rise and widespread adoption of bibliometrics have relied on the development of computer-based indexes and databases that enable the capture of the necessary bibliographic metadata and allow that metadata to be stored, linked, searched, shared, and ultimately analyzed using mathematical methods.

What Is Bibliometric Data? Bibliographic Metadata as the Input to Bibliometrics

As suggested by the history and definition of bibliometrics, the core of bibliometric data is based on the bibliographic metadata available about a wide range of document types. Today, bibliometric data largely relies on indexing and citation databases that capture an ever-expanding and robust set of bibliographic metadata and are therefore also constrained by it. Table 1.1 lists the most common bibliographic metadata that underpins most bibliometric calculations. There is often confusion between what is considered bibliographic data and what is considered bibliometric data. Although there is certainly overlap, data transitions from bibliographic data into bibliometrics when it is aggregated, counted, or used in some mathematical formula. In other words, bibliographic data is the input, and bibliometrics is the output.

Something should also be said about the types of documents that are included in bibliometric analysis. Since bibliometrics is dependent on the databases that capture the needed data, the types of publications available in these databases dictate what can be included. Typically, indexing and citation databases (such as Web of Science, Scopus, etc.) include these publication types: journal articles, review articles, conference proceedings papers, books, book chapters, editorials, and letters. Some also capture errata, corrections, and preprints. However, there are also an increasing number of new document types being

Table 1.1: Bibliographic metadata used in bibliometrics

Bibliographic Metadata Types	Metadata Fields
Document	Document title Journal/book/source title Publication year Volume/issue Page numbers ISSN/ISBN Document level identifiers (DOI, PMID, ArxivID, etc.) Publication type Source type Language Open access status
Author	Author names Author identifiers Affiliation name Affiliation address Affiliation country
Content	Abstract Author keywords Indexed keywords Journal level classifications Article level classifications
Citation	Reference list Document level citation count
Funding	Funding body name Funding body address

captured, including data papers and short surveys.[1] The set of publication types is ever-expanding as institutions and researchers recognize the biases in bibliometrics when it focuses on only a discrete set of publication types.

More generally, almost any metadata field that is supplied about a document has the potential to be added into a bibliographic data set; therefore, bibliometric systems or tools are continually adding to the complexity of the available data.

What Are the Major Sources of Bibliometric Data?

There are surprisingly few data sources of bibliographic metadata that include the full suite of data required to perform robust bibliometrics. The challenge lies in providing extensive citation linking between the source document, reference lists, and those that cite it. This exponentially expands the size of the data set and poses an insurmountable challenge (or at least a pretty large challenge) for many bibliographic databases. The network (or citation mapping, as it is often called) created through this linking is the fundamental power of the data sources used for bibliometric analysis.

To date, only a few bibliometric sources provide citation linking within their bibliographic data. These include the following (Visser, van Eck, and Waltman 2021):

- **Web of Science:** Owned by Clarivate Analytics, Web of Science is a very large multidisciplinary database that is made up of several indices to which an institution can subscribe selectively. These indices are made up of a curated list of journals and publications that are reviewed against quality standards for inclusion. Although the subject areas covered are still heavily focused on the sciences, Web of Science continues to grow its coverage of social sciences, arts, and humanities.
- **Scopus:** Owned by Elsevier, Scopus is a very large multidisciplinary database that is largely made up of a curated list of journals and publications, which are reviewed against quality standards for inclusion. The subject areas covered are still heavily focused on the sciences, with approximately 27 percent (as of April 2022) of its content on the social sciences (which include arts, humanities, business, economics, decision sciences, and psychology; Scopus 2022).
- **Dimensions:** Owned by Digital Science, Dimensions is a very large multidisciplinary database that ingests metadata from freely available online sources such as Crossref, PubMed, and PubMed Central and then supplements the database with licensed content directly from publishers. The Dimensions platform is also a bibliometric assessment tool, making it different from Web of Science and Scopus, which offer primarily the bibliographic source data with limited analytical tools. Dimensions also provides some free access to its system and noncommercial access to its data via API.
- **Crossref:** Owned by Publishers International Linking Association, Crossref is a not-for-profit metadata service that allows its members—made up of publishers, institutions, funding agencies, and government agencies—to register their content and mint DOIs for the purposes of reference linking. It provides free access to their metadata and encourages ingest into external systems for wide use.
- **OpenAlex:** As a response to Microsoft Academic pulling the plug in late 2021 (a huge blow to open-source systems engineers; Singh Chawla 2021; Microsoft 2021), the nonprofit company OurResearch developed OpenAlex. It adheres to its open-source principles making its index of research entities—such as scholarly papers, authors, and institutions—available openly on its web application via API and an entire local database download snapshot for offline access.

There is a difference between the bibliographic sources listed above and the bibliometric tools used to analyze the data. This section focuses on the sources

of bibliographic data that feed into the bibliometric tools that will be further explored in chapter 2. However, for many of these data sources and bibliometric tools, the lines are becoming blurred as more data sources are integrated into existing systems and as new companies emerge and form new innovative tools and integrate existing ones.

There is also a growing number of systems that provide robust bibliometric data but are not considered sources of bibliometric data because they are not primarily used to feed into external systems for additional analysis. Systems such as Lens.org (Cambia), the Leiden Ranking (Centre for Science and Technology Studies), and Scite.ai use external data sources, then supplement the data with in-house reference linking intended to enhance their citation analysis capabilities.

Although many bibliometric tools are available, nearly all draw on at least one of the bibliographic sources Web of Science, Scopus, Crossref, or OpenAlex (and previously Microsoft Academic). However, the distinction between the data sources for bibliometric systems is further blurred as new, highly intelligent multipurpose systems emerge, such as Scite.ai, Lens.org, and Bibliometrix (see chapter 2 for even more systems), that integrate ever-expanding types of data sources. Bibliometric tools no longer simply produce bibliometrics based on citation linking between scholarly documents such as journal articles, conference proceedings, or books. They now provide metadata on grants, patents, clinical trials, research data sets, policy documents, and more. An increasing number of tools are linking this complex data together from across content types to provide more complete profiling opportunities at the institution, department, and researcher levels. For example, many of the tools link patents and articles to provide a count of articles that have received patent citations (see chapter 3 for a use case). These articles can then be analyzed by a number of factors, such as research areas, top author, coauthorship, geographic distribution, and so on.

Google Scholar has yet to be mentioned. This is primarily due to its virtual lack of usefulness as a data source. Google Scholar provides a massive searchable database of scholarly materials, including citation linking and citation counts that often outnumber other systems; however, it does not allow the data to be easily exported to other systems or linked to from other systems for use, thereby disqualifying it as a practical bibliometric data source. With that said, Harzing's Publish or Perish (PoP) does pull Google Scholar data into its system (using an API access token) and offers a number of citation indicators. Still, PoP has to deal with the annoying issue of Google Scholar sending CAPTCHAs when the PoP system sends queries at too high a rate, just to make sure there is an actual person using the PoP and it is not just a massive data harvester (Harzing.com 2022). This roadblock is likely why every other system out there does not attempt to ingest Google Scholar data.

What Are Bibliometric Indicators?

Bibliometric indicators are the output of bibliometrics built from bibliographic metadata. Indicators make visible otherwise invisible phenomena. Bibliometrics is intended to answer questions about research productivity, impact, excellence, collaboration, networks, and more. However, these phenomena are largely unobservable unless proxies are used to represent them. Therefore, observable measures—such as counts of authors, documents, citations, affiliation, and so on—are used to represent the unobservable phenomena to be examined (Sugimoto and Larivière 2018). Cassidy Sugimoto and Vincent Larivière, as well as Yves Gingras (Gingras 2016), offer excellent further reading on the validity of indicators, which is beyond the scope of this report.

Bibliometric indicators are typically divided into several groups:

- **Productivity indicators** give insight into the activity of an entity,[2] measured through publication/document counts.
- **Impact indicators** give insight into the attention given to a document or set of documents of an entity, usually measured through citation counts, including citations in other research output, policy, clinical trials, knowledge syntheses, or patents.
- **Collaboration indicators** give insight into the amount of overlap between two entities and the nature of this overlap, usually measured by coauthorship and affiliation data.
- **Interdisciplinarity indicators** give insight into the overlap of research areas by looking at research outputs resulting in collaborations between authors from different disciplines or at research outputs citing or spanning more than one discipline.
- **Alternative metrics** are a growing set of indicators that may eventually evolve into separate standard categories on their own. However, this report groups them together not because of their newness but because the selections of indicators, their data sources, and the ways these indicators are grouped and presented are so different among the various bibliometric tools that it is difficult to describe with any consistency. However, in general, alternative metrics include the following indicators:
  - social media attention, measured through tweets, likes, blog post links, and so on
  - views and use, measured through database/

Table 1.2: Introductory bibliometric indicators

Productivity Indicators			
Metric	***Description***	***Variations***	***Caveats***
Number of documents	The total number of documents from a specific entity (e.g., country, institution, group, etc.).	Temporal analysis options are common (e.g., number of documents per year). The selection of publication types included will impact counts.	Avoid making comparisons across disciplines, author at different career stages, and entity size. Suggest using productivity alongside other normalizing indicators and/or with trend information. The data source determines the publications that are included.
Impact Indicators			
Metric	***Description***	***Variations***	***Caveats***
Number of citations	Total number of citations received from a subset of documents (e.g., country, institution, group, etc.).	Include or exclude self-citations. Percent of the total number of documents.	Expected to trend downward toward most current years. Citations take time to accumulate. Avoid making comparisons across disciplines, author at different career stages, and entity size.
Number of documents cited	Total number of documents that have received at least 1 citation.	Include or exclude self-citations. Percent of the total number of documents.	Avoid making comparisons across disciplines, author at different career stages, and entity size.
Number of citations per paper	Total number of citations divided by the number of documents within an entity (e.g., country, institution, group, etc.).	Include or exclude self-citations.	Expected to trend downward toward most current years. Citations take time to accumulate. Avoid making comparisons across disciplines, author at different career stages, and entity size.
Number of documents in the topmost cited documents worldwide	Calculates the total number of documents from an entity (e.g., country, institution, group, etc.) that are in the top percentages of all cited documents.	Most often available as top 1%, 10%. Percent of the total number of documents. Can be field- or subject-weighted. Temporal analysis options are common. Include or exclude self-citations.	Can be a relatively small number causing large variances in trend data.
Normalized citation impact	Normalization usually occurs by discipline/subject/field, publication year, and publication type. The normalized value is then compared to an expected normalized global value, and the metric is represented as an index relative to 1 that indicates the expected global value.	Include or exclude self-citations. Journal normalization uses the journal the document is published in for normalization rather than the subject, year, and publication type.	Small entities, such as individual authors or small groups, will see large variances in trend data.
Collaboration Indicators			
Metric	***Description***	***Variations***	***Caveats***
Number of documents with coauthor at another institution	Total number of documents with coauthors at another institution or entity type.	Percent of the total number of documents. Corporate or industry/academic coauthorship. National/domestic coauthorship. International coauthorship. Institutional coauthorship.	Collaboration practices vary among disciplines, institution type, and geographic regions.

Table 1.3: Why leaders should care about responsible metrics

Arguments	Description
Maintain institutional autonomy	Take control of what is actually being measured to align with your institution's values, as opposed to being reactive to external evaluations.
Make better decisions	Ensure that what you intend to evaluate is really being measured, aligning the indicators with the phenomena of interest. Make sure bibliometrics is truly needed for the evaluation, and include other, qualitative approaches.
Ensure return on investment	Ask whether the bibliometrics approach is the best way, and if it is, ensure that there is sufficient investment to pursue meaningful evaluation.
Establish operational readiness	Understand what responsible research evaluation policies and external expectations exist or are in development.
Manage reputational risk and enhance staff well-being	Get ahead of poorly used metrics, especially where they impact individual reputations and well-being. Too much attention to metrics begets too much attention to metrics. Balance in the evaluation and assessment should include a compassionate approach to the impact on scholars' work and life.

Source: Adapted from the INORMS Research Evaluation Group documentation. For full descriptions, see INORMS Research Evaluation Group 2020a.

reference manager clicks, downloads, and saves
- media attention, measured through news media coverage
- recommendations, measured through post-publication peer review

There is a wide variety of bibliometric indicators, too many to provide a comprehensive overview here. However, table 1.2 provides an overview of the main indicators commonly presented in bibliometric assessment tools. These are largely descriptive analyses, which must be distinguished from more rigorous statistical analyses. Ana Andrés's book *Measuring Academic Research: How to Undertake a Bibliometric Study* provides excellent in-depth guidance on selecting specific indicators to apply to a bibliometric research study (Andrés 2009). However, for more detail, see chapter 3 of this report, which will provide some guidance on general applications of bibliometric indicators.

The indicators presented in table 1.2 are derived in large part from the bibliometric tools SciVal (Elsevier 2019), Incites (Clarivate 2022), Dimensions (Dimensions 2021), and Lens.org (Lens.org 2022). These tools are pervasive within the global bibliometric community and provide a bridge between bibliometric practitioners, who typically provide bibliometric services to their institutions and academic staff with an entry or intermediate level of experience with bibliometrics, and expert-level bibliometric services or researchers, who conduct more advanced calculations and methodologies that require data science or statistical knowledge (Lancho Barrantes, Vanhaverbeke, and Dobre 2021; Cox et al. 2019). This is an important distinction, as bibliometrics is a complex historied field, and practitioners (often librarians within academia) can feel adrift when faced with the mathematical complexities of the statistical analyses presented in bibliometric research papers. In addition to these tools, the University of Waterloo's white paper "Measuring Research Output through Bibliometrics" was also used as a guiding information source (Byl et al. 2016).

Responsible Use of Metrics

A highly effective bibliometric practitioner is . . . value-led, not data-led.

—Gadd 2020

It is essential that bibliometrics be approached with a duty to responsible use. This becomes particularly salient when working with individuals holding decision-making power at all levels throughout academia. The INORMS Research Evaluation Group (2020a) outlines several reasons senior leaders should be interested in responsible evaluation of research, which include maintaining institutional autonomy, making better decisions, ensuring return on investment, establishing operational readiness, and managing reputational risk and enhancing staff well-being (table 1.3). The work coming out of INORMS is crucial for a bibliometric practitioner to understand. Its work is a response to the vast amount of professional experience of its members and integrates learnings from the impacts of the UK Research Evaluation Framework and leading internationally recognized statements of principles such as the Leiden Manifesto (Hicks et al. 2015), the San Francisco Declaration of Research Assessment (DORA 2012), and the Metric Tide report (Wilsdon et al. 2015; see also table 1.4). Therefore, the resulting INORM *SCOPE Framework* (INORMS Research Evaluation Group 2020b) is intended to be a bridge between the ideals stated in these principles and the business of doing bibliometrics.

SCOPE is an acronym for Start, Context, Options, Probe, and Evaluate, and the *SCOPE Framework* is

Table 1.4: Snapshot of responsible use of metrics declarations, manifestos, and principles

	Milestones	Audience	Key Messages
DORA—"San Francisco Declaration on Research Assessment" (DORA 2012)	established 2012 22,162 signatories as of Sept. 2022	funding agencies institutions publishers metrics suppliers researchers	Eliminate journal-based metrics that mask the merits of individual works (aka dump the JIF). Release online publications from the confines of print.
The Metric Tide report (Wilsdon et al. 2015)	released 2015 revisited 2022	anyone conducting quantitative evaluations of research particularly relevant to the UK Research Excellence Framework environment	Set of five principles for the use of any metrics. Considers robustness, humility, transparency, diversity, and reflexivity in the use of any quantitative indicators.
"The Leiden Manifesto" (Hicks et al. 2015)	released 2015	administrators practitioners researchers	Set of 10 principles for the use of bibliometrics. Quantitative evaluations should not supplant qualitative ones. If they are used, however, they should be situated appropriately within the context of the researcher, the institution, the region, and the field. Transparency is key.
INORMS SCOPE Framework (INORMS Research Evaluation Group 2020a)	released 2021	anyone conducting evaluations of research	Start with value. Make evaluations context-specific. Consider the validity of the methods. Probe deeper into latent consequences. Evaluate the evaluations.

meant to guide the practitioner through a series of steps that help with adherence to responsible use of metric principles. These steps walk through fundamental questions helping the practitioner to first question the reasons for using metrics, match the level of analysis to the need, identify appropriate methodologies, dig deeper into the potential and known effects of the analysis, and finally evaluate whether the goals of the analysis were achieved. Following these steps is practical, and beyond setting up any bibliometric project with sound judgments, they can also be used as a communication tool for establishing standard expectations for research evaluation within institutions.

Notes

1. These are new document types from Scopus (Elsevier 2020): data papers are short descriptive papers about data sets; errata report errors, corrections, and retractions of published papers; and short surveys are short (only a few pages) reviews of research.
2. Entities can represent a person, group, institution, region, etc., and can be subdivided by subject, discipline, years, etc.

References

Andrés, Ana. 2009. *Measuring Academic Research: How to Undertake a Bibliometric Study.* Amsterdam, The Netherlands: Elsevier.

Byl, Lauren, Jana Carson, Annamaria Feltracco, Susie Gooch, Shannon Gordon, Tim Kenyon, Bruce Muirhead, et al. 2016. "Measuring Research Outputs through Bibliometrics." Working paper. January 2016. https://doi.org/10.13140/RG.2.1.3302.5680.

Clarivate. 2022. "Normalized Indicators." InCites Help. Accessed September 22. https://incites.help.clarivate.com/Content/Indicators-Handbook/ih-normalized-indicators.htm.

Cox, Andrew, Elizabeth Gadd, Sabrina Petersohn, and Laura Sbaffi. 2019. "Competencies for Bibliometrics." *Journal of Librarianship and Information Science* 51, no. 3: 746–62. https://doi.org/10.1177/0961000617728111.

Dimensions. 2021. "Which Indicators Are Used in Dimensions, and How Can These Be Viewed?" Last modified September 19. https://dimensions.freshdesk.com/support/solutions/articles/23000018839-which-indicators-are-used-in-dimensions-and-how-can-these-be-viewed-.

DORA. 2012. "San Francisco Declaration on Research Assessment." December 15, 2012. https://sfdora.org/read/.

Elsevier. 2019. *Research Metrics Guidebook*. https://elsevier.widen.net/s/chpzk57rqk/acad_rl_elsevierresearchmetricsbook_web.

Elsevier. 2020. *Scopus Content Coverage Guide*. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf.

Gadd, Elizabeth. 2020. "The Five Habits of Highly Effective Bibliometric Practitioners." Keynote address, Bibliometrics and Research Assessment Symposium 2020, online, October 7–9. National Institutes of Health Library. YouTube video, 58:01. https://youtu.be/SFi1NbbSdQ8.

Garfield, Eugene. 1955. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas." *Science* 122, no. 3159: 108–11.

———. 1972. "Citation Analysis as a Tool in Journal Evaluation." *Science* 178, no. 4060: 471–79. https://doi.org/10.1126/science.178.4060.471.

Gingras, Yves. 2016. *Bibliometrics and Research Evaluation: Uses and Abuses*. Cambridge, MA: MIT Press.

Harzing.com. 2022. "Why Do You Give Me These Stupid CAPTCHAs?" Frequently Asked Questions, last updated June 2. https://harzing.com/resources/publish-or-perish/manual/about/faq#004.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. "Bibliometrics: The Leiden Manifesto for Research Metrics." *Nature* 520, no. 7548: 429–31. https://doi.org/10.1038/520429a.

INORMS Research Evaluation Group. 2020a. "Five Arguments to Persuade HE Leaders to Evaluate Research Responsibility." https://inorms.net/wp-content/uploads/2020/06/five-ways-to-persuade-leaders-to-evaluate-responsibly.pdf.

INORMS Research Evaluation Group. 2020b. *The SCOPE Framework: A Five-Stage Process for Evaluating Research Responsibly*. INORMS Research Evaluation Group. https://inorms.net/wp-content/uploads/2021/11/21655-scope-guide-v9-1636013361_cc-by.pdf.

Lancho Barrantes, Barbara S., Hannalore Vanhaverbeke, and Silvia Dobre. 2021. "2021 Competency Model for Bibliometric Work." *Bibliomagician* (blog). https://thebibliomagician.wordpress.com/competencies/.

Lens.org. 2022. "Glossary." Support. Accessed September 22. https://support.lens.org/glossary/.

Microsoft. 2021. "Next Steps for Microsoft Academic—Expanding into New Horizons." *Microsoft Academic* (blog), May 4. https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/.

Mokhnacheva, Yu. V., and V. A. Tsvetkova. 2020. "Development of Bibliometrics as a Scientific Field." *Scientific and Technical Information Processing* 47, no. 3: 158–63. https://doi.org/10.3103/S0147688220030036X.

Pritchard, A. 1969. "Statistical Bibliography or Bibliometrics?" *Journal of Documentation* 25, no. 4: 348–49.

Scopus. 2022. "Scopus Sources List." https://www.scopus.com/sources.

Singh Chawla, Dalmeet. 2021. "Microsoft Academic Graph Is Being Discontinued. What's Next?" News, Nature Index. June 15. https://www.nature.com/nature-index/news-blog/microsoft-academic-graph-discontinued-whats-next.

Sugimoto, Cassidy R, and Vincent Larivière. 2018. *Measuring Research: What Everyone Needs to Know*. Oxford: Oxford University Press.

Visser, Martijn, Nees Jan van Eck, and Ludo Waltman. 2021. "Large-Scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic." *Quantitative Science Studies* 2, no. 1: 20–41. https://doi.org/10.1162/qss_a_00112.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, et al. 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. https://doi.org/10.13140/RG.2.1.4929.1363.