Bibliometrics in the context of research evaluation and research policy

Purnell, P.J.

Chapter 1

General introduction

1.1 Overview of the dissertation

Bibliometrics is the quantitative measurement of various aspects of research including productivity, impact, subject categorisation, science networks, and time trends. Such bibliometric measurements are made based on data about scientific publications. Key consumers of bibliometric reports are research policymakers and those conducting research evaluation. Bibliometrics is a useful tool because it is relatively quick, cheap, and easy to use, although research evaluators and policymakers may have other tools available, especially the use of expert opinion, which is more resource-intensive but conducted by peers with subject matter expertise. There is a lot of debate in the research community about how bibliometrics and expert opinion should be combined, or whether one approach is better than the other. The aim of this dissertation is to demonstrate the benefits of bibliometrics to policymakers and other stakeholders, while highlighting and analysing the causes of its limits. Research evaluators and policymakers will benefit from a good understanding of the advantages and limitations of bibliometrics and how expert opinion can enhance the big picture presented by bibliometric analysis.

The following sections of chapter 1 comprise an introduction to research policy and research evaluation, followed by a detailed history and description of bibliometric techniques used in research policy development and research assessment. The chapter concludes with a summary of the motivation and objectives for the dissertation. Chapters 2 – 6 comprise five bibliometric studies that illustrate how quantitative techniques can generate large-scale database comparisons, subject maps, trend and geographical analyses that can inform research policy and evaluation. These studies also highlight some of the limitations of bibliometrics related to data reliability and the need for complementary expert judgement. Across the studies, some common themes emerge that may improve the usefulness of bibliometrics through community-led initiatives and new technologies.

Chapter 7 provides a summary of the findings of each of the studies along with the overall implications. It becomes clear that bibliometrics is not only measuring, but also shaping the research system. Lessons are learned on how bibliometrics can have both desired and unintentional effects on the research system. Key takeaways from the studies on the debate between bibliometrics and expert opinion result in recommendations on finding the ideal balance between these approaches. Positive effects of community-led initiatives already noticeable in bibliometrics are encouraging and suggest there are more ways the community can build on these gains. The future of bibliometrics in research evaluation and policymaking will depend on improvements in data quality and curation, and on ensuring it shapes the research community in a positive way.

1.2 Research policy

Many organisations and regional bodies offer differing definitions of research policy, science policy, and a range of related policies. Science policy can be seen as making public funds ‘more effective and more efficient’ at new knowledge creation (Organisation for Economic Co-operation and Development, 2024). The European Commission’s research & innovation policy aims to address major societal challenges by ‘converting them into opportunities’ for innovation (DG Research and Innovation, 2020). The Commission holds an annual research & innovation policy forum known as the European Research & Innovation Days that aim to convene policymakers, citizens, and other stakeholders to co-design its action plan (European Commission, 2024). The United States science and innovation policy aims to

‘strengthen research and innovation enterprise’ in pursuit of national goals (National Academies, 2024). National or regional science policies often feed into the strategic plans of individual institutions, which may develop their own research policy. Institutional research policies are often concerned with aspects of research that include research ethics, equity, diversity, and inclusivity, open science, and intellectual property, and are closely related to the distribution of research funds (e.g. Cambridge University, 2024; New York University, 2024; University of California Berkeley, 2024; University of Pretoria, 2023; University of Technology Sydney, 2024).

The definitions of science, research, innovation, and other related policies are therefore not clearly distinguished or consistently defined and definitions often overlap, but research policy can generally be seen as a component of a broader science and innovation policy. For the purposes of this dissertation, I consider science policy as a broader term that governs the overall national approach to innovation, while research policy is more concerned with activities conducted at institutional or departmental level. Although the topics discussed here address challenges at a range of levels, I will from here on refer only to research policy for the sake of clarity.

Research policymakers are usually involved in governance or allocation of financial resources to be spent on research. Recipients of research funding may include universities, research institutions, private industry, nonprofit organisations, individual researchers, postgraduate students, or government entities. Typically research grants are used to pay salaries, buy or loan the use of equipment, build new laboratories, cover travel for members of the research team, and pay the costs of publishing research findings. Research policymakers therefore sometimes make use of bibliometric analyses to inform policy development.

In the modern world, scientific knowledge has become a resource upon which societies are built, technological advances based, and for which people compete as they do for other commodities. In this competitive environment, approval of research projects, hiring of academics, and assignation of research funds require justification and are frequently subject to a selective evaluation process. Techniques to assess research productivity, performance, and impact including bibliometrics have therefore been developed to support decision-making and have become a key component of research policy.

1.3 Research assessment

There are two broad approaches to research evaluation, one is to seek expert opinion of researchers familiar with the research being assessed in processes collectively known as peer review. The other is to aggregate quantitative performance indicators such as impact factors and is described as bibliometrics. There are strong arguments for and against both peer review and bibliometrics in research evaluation with research policymakers having to carefully set out their assessment procedures by combining elements of both.

In the next part of this introduction, I will give an overview of the arguments in favour of and against both peer review and bibliometric techniques and then introduce the concept of a middle path which incorporates elements of both methods. Thereafter, I will concentrate on bibliometrics and its relationship with various research policies, which is the subject of this dissertation. Bibliometrics is sometimes seen simply as a tool to support research assessment. I take a broader view that bibliometric

techniques go beyond assessment and can be used for other types of studies such as describing the structure of fields, collaboration network and trend analysis, and science mapping studies. I show that bibliometrics can sometimes mislead us and raise questions about the extent to which we are comfortable relying on bibliometric techniques and when we should incorporate a heavier emphasis on peer review and expert judgement more in general.

1.3.1 Peer review

Since 1731, members of the Royal Society of Edinburgh have been asked to review contributions prior to publication, a process followed by the Royal Society of London in 1752 (Spier, 2002). Since then, the academic community has judged research quality among fellow scholars, asking peers to review each other's scientific manuscripts and give their professional opinion on research proposals, applications for academic posts, and many other processes that require specialised knowledge. Peers' expert opinion is intended to provide a unique perspective based on deep understanding of the topic, scientist, or process being assessed.

With this unique perspective comes the risk of prejudiced personal judgement where the reviewer is influenced by preferences, or fear of negative perception and consequences of the review. To encourage candid review, the reviewer's identity is often masked (single-anonymous review) and to mitigate potential reviewer bias the identity of those being reviewed is frequently anonymised (double-anonymous review).

Conducting, writing, and publishing academic research requires knowledge and adherence to specific scientific protocols. Other academics are therefore uniquely placed to judge the quality of another scholar's contribution. The peer review process has been shown to improve manuscripts through suggestions on presentation, readability, and through identifying missing references and scientific or statistical errors (Ware, 2008). Indeed, the very fact that manuscripts are subjected to the peer review process is likely to incentivise authors to invest effort to maximise the chance of their manuscript being favourably judged and accepted for publication (Ware, 2008) and can even improve reviewers' own writing (Lundstrom & Baker, 2009).

Critics of the peer review system point to a number of problems. One common argument against peer review concerns the identity of the reviewer or reviewers. Surely the selection of review committee members will determine the outcome of the review. If alternative members were appointed to the committee, the result would likely be rather different (Bertocchi et al., 2015). Another argument surrounds the risk of bias in which reviewers make prejudiced decisions based on their perceptions of the people or institutions they are evaluating. This introduces the risk of creating self-serving networks that support those scientists who are already in the club and make it difficult for newcomers to gain access or approval of the network. Young people or those newly arrived from other places who have not yet made their mark could inadvertently miss out on positive acclaim in favour of colleagues with established names. Another criticism of peer review centres around the costs involved and the burden on the reviewers, especially in the context of national level research assessments (Geuna & Martin, 2003; Martin & Whitley, 2010).

In summary, the processes behind qualitative evaluation of researchers have been described as resource-intensive, and error-prone (Ioannidis & Maniadis, 2023). The peer review process offers benefits to the

scientific process but clearly requires a complementary, objective process to provide balanced research assessment. For a full review of the peer review process, see Lee et al. (2013).

1.3.2 Bibliometric approaches

The existence of large amounts of stored data associated with published scholarly works and the availability of analytical tools means that policy makers can quickly and easily observe the state of certain aspects of science and detect trends that will enable informed decision making. The increase in publication data and associated needs for new mechanisms to manage it has been described as the industrialisation of research (de Solla Price, 1978).

The use of bibliometric analysis in the evaluation of science relies on the assumption that production of new knowledge and the recognition of peers are the pillars of scientific impact (Desrochers et al., 2018). In order to measure new knowledge production, research outputs such as academic papers (and also other outputs including books, patents, and datasets) can be counted, so that a greater number of papers is an indicator of higher scientific productivity. Scientific authorship has been described as the 'undisputed coin in the realm of academia' (Cronin, 2001) and is routinely used to assess productivity of researchers, departments, universities, and countries. The link between the author and institution (and therefore country) is made by the affiliation mentioned on the manuscript submitted to the publisher and facilitates analysis of institutional and international collaboration networks and enables rankings of scientists, universities, and countries.

Authors of studies that quote from, or use the ideas presented in earlier papers are expected to refer to the original source in the cited references at the end of their paper. Citations have therefore been described as 'pellets of peer recognition' (Merton, 1988). In other words, authors cite other scholarly works because the cited work has in some way influenced the author in the writing of their own paper (Merton, 1973). The author acknowledges this influence in the form of a cited reference, which allows the author to use the cited work without committing plagiary (Merton, 1988) and which forms the basis of the incremental nature of scientific discovery. For an in-depth review of the reward system of science, see e.g., Desrochers et al. (2018). Large-scale aggregation of publications and citations forms the basis of bibliometric analysis and serves as a proxy for measuring scientific productivity and impact. Indeed, quantitative bibliometric analysis presents policy makers with a quick, easy, and cost-effective complement to peer review.

Some of the assumptions described above do not always hold true and have been used to question the validity and reliability of bibliometric techniques in research evaluation. Authorship can be complicated by questions about the distinction between contributions that warrant full authorship versus those that merit only acknowledgement (Costas & van Leeuwen, 2012; Desrochers et al., 2017), the hierarchical order in which authors are listed, and the decline in single-author papers (Nabout et al., 2015), among others. The motivation of authors to cite each others' work is also the subject of much debate (Knorr-Cetina, 2013; Nicolaisen, 2007). Problems include the potential for manipulation through citing one's own work or that of colleagues to manipulate citation metrics (Davis, 2012; Fister et al., 2016).

Practical matters also need to be considered when designing or interpreting bibliometric studies. For instance, the usual source of publication and citation data for analysis is a citation database. Any research outputs not indexed in the source database used cannot be included in the analysis. Database

coverage therefore becomes one of the defining factors in the outcome of any bibliometric study. We will examine the main bibliometric data sources in section 1.3.

1.3.3 The middle path

As there is no ground truth that defines research quality, we cannot judge whether the peer review or bibliometric approach is more accurate. It may be more prudent to judge the extent to which the two approaches reach consensus, that is to what extent do peer review and bibliometric analysis agree? Some studies have shown broad agreement between the results of peer review and bibliometric findings (Bornmann & Leydesdorff, 2013; Pride & Knoth, 2018; A. Van Raan, 2006), while others have shown weaker correlation (Mryglod et al., 2015b, 2015a; Wilsdon et al., 2015). There appears to be substantial variance in correlation between peer review and bibliometrics depending on the level of aggregation at which the comparison is made (Traag & Waltman, 2019), whether size-dependent or size-independent correlations were used (Mryglod et al., 2013b, 2013a), and when assessing really high-quality research (Bertocchi et al., 2015; CWTS Leiden University, 2007; Rinia et al., 1998).

Although peer review and bibliometrics are often discussed as separate approaches, they are in fact intricately linked and each contains a component of the other. Expert opinion may be influenced by bibliometric indicators and bibliometric studies are often aggregates of expert opinion. For example, several major ranking systems ask academics to vote on the prestige of universities which, when combined with publication and citation counts, determines the rank of universities. On the other hand, when an academic cites another author's papers in their article, this is usually seen as an intellectual endorsement of the paper based on the academic's expert opinion. By counting citations then, we are summing up expert opinions. This shows that there is no clear-cut separation between bibliometric indicators and expert opinion. In practice one will almost always be using a combination of input derived from bibliometric indicators and expert opinion.

Nevertheless, opinions supporting the cases for peer review and bibliometric analysis while eschewing the other are strong, and it can be tempting to see peer review and bibliometric analysis as two opposing camps to choose between. It is easy to take an extreme position that points to the benefits of one camp while emphasising drawbacks of the other. However, embracing the benefits of bibliometric indicators and peer review while taking steps to address their drawbacks is likely to provide the most agreeable way forward in research assessment.

Policy makers and other research stakeholders eagerly seek clarity on the right path to follow; should we embrace, reject or improve bibliometric methods? In this dissertation, I opt for the path of improvement. It is a rocky path, strewn with obstacles and open to criticism from both extremes of the argument. Some of the criticism is valid, but as argued above, thinking in a simplistic way about rejecting bibliometrics in favour of peer review, or the other way around, is counterproductive and misses the crucial point that the two are intricately linked. To find balance, we need a more nuanced perspective, which is what this dissertation aims to offer. Such a perspective is sorely needed to find a middle path as the basis for pragmatic use of bibliometrics by policymakers and evaluators. By identifying weaknesses in the bibliometric system, we can contribute to ways to address these weaknesses and thereby improve the system. This will make the middle path that incorporates elements of both peer review and bibliometric analysis smoother to travel.

1.4 Bibliometric analysis

In this section, I will provide an overview of how bibliometric analysis began and developed into an integral part of systemic research policy development and research evaluation. I will describe the impact of key moments such as the launch of the first citation index, changes brought by the World Wide Web, and the mass meeting of academic minds that insisted on responsible use of metrics. I will also use the section to present a summary of the major data sources used in bibliometric analysis and the important differences between them. These differences are highly relevant because choosing one or another data source can significantly influence the result of any bibliometric analysis. I will also cover the main indicators used in bibliometrics and describe what they are supposed to measure. Due to the popularity (or notoriety) of university rankings and because of their dependence on bibliometric analyses, I will present the major international ranking systems. Finally, I will briefly discuss science mapping. As the amount of data to be analysed has grown so large, creation of science maps has proven a useful way to easily visualise patterns and trends in the networks. I will present a map of the terms that most frequently occur in the five articles of this dissertation to visualise the connections between them.

1.4.1 A brief history of bibliometrics

In one of the first bibliometric studies, F.J. Cole, a professor of zoology and Nellie B. Eales, a museum curator, counted the number of research papers on comparative anatomy published by authors in European countries between 1543 and 1860 (Cole & Eales, 1917). Originally intended to be a study of anatomical museums, the authors realised that the number of museums (537) was too small to draw statistical conclusions. They switched their attention to scholarly literature because research papers are permanent, accessible and it is easy to define who conducted the study and when (Cole & Eales, 1917). At about the same time, there was a shift in higher education from the general cultural education provided by the small community college towards the demand for the expert worker equipped with specialised knowledge acquired from large universities that shifted emphasis towards graduate education (Gross & Gross, 1927). Librarians had to quickly learn how to develop collections of relevant scientific periodicals to cater for their specialist departments and began using citation counts as a method of assessing the relative value of journals when building and maintaining their collections. Librarians at Pomona College in Claremont, California listed the number of references from a single volume of the Journal of the American Chemical Society to other periodicals in successive five-year periods and then ranked the cited journals in order of relative importance to the field (Gross & Gross, 1927).

Following these early works, historians continued using publication analysis to further their understanding of their field and librarians developed techniques to maximise the usefulness of their collections. However, it was the advent of 'big science' that led to the shift from library-based 'statistical bibliography' to the large-scale use of bibliometrics as a key tool for the policy maker (Narin & others, 1976). The need to constitute a coordinated set of measures related to books and documents was first described in French as 'La bibliométrie' (bibliometrics) by Paul Otlet in his *Traitée de Documentation. Le livre sur le Livre. Théorie et Pratique* (Otlet, 1934), and later Alan Pritchard's definition of the academic field of bibliometrics as 'applying mathematical and statistical methods to books and other media of communication' (Lawani, 1980; Pritchard, 1969) was adopted in English (Hood & Wilson, 2001). The closely related term 'scientometrics' is generally accepted to more broadly

encompass measurement of any aspect of science and technology and originated from Vassily Nalimov's use of the Russian 'Naukometriya' (Nalimov & Mulchenko, 1969).

A defining moment came in 1955 when Eugene Garfield published his work showing how lists of cited references from scholarly papers could be used to create a network of scholarly publications in specific fields linked by citations (Garfield, 1955). Garfield then founded the Institute of Scientific Information (ISI) in Philadelphia and promptly launched the Science Citation Index. Counting up the citations to articles published in specific journals was considered a proxy for the journal's influence and became a counterweight to the opinions of science leaders. The Journal Impact Factor (JIF) was developed throughout the 1960s (Garfield, 1972), commercialised in 1975, and has played an important role in countless decisions by policymakers (Larivière & Sugimoto, 2019), although Garfield did not personally endorse its use as a performance indicator (Wouters, 2017). One argument against widespread use of the JIF as a performance indicator was the concern that the metrics weren't entirely accurate (Moed & Van Leeuwen, 1995) due to widely varying citation dynamics between journals for certain document types.

In an increasingly data-driven world, various national and multinational bodies began to turn their focus toward measuring and analysing progress of science and technology. The U.S. National Science Board's *Science Indicators* report, first published in 1973 (National Science Board, 1973), presented a system of science indicators to 'describe the state of the scientific endeavour'. According to the report, the indicators were intended to be adapted over time and become the basis for discovering trends and identifying areas of strength and weakness.

Across the Atlantic, the Europeans were keen to investigate potential applications of bibliometric analysis in strategic decision making. In 1980, Leiden University conducted the first detailed bibliometric analysis of 140 research groups in response to a decision by the University Executive Board to allocate funding partially based on research performance. The quantitative analyses were complemented by interviews with the researchers being studied and supported by peer review. Encouraged by this study, the university executive board supported the development of a group working on new and improved bibliometric techniques, which a decade later was formally named the Centre for Science and Technology Studies (CWTS) (van Raan, 2021).

There followed a rush to develop new bibliometric indicators for use in large-scale assessments of scientific communities to support policy making at various levels in many countries. Taking the Netherlands as an example, the Association of Universities in the Netherlands (VSNU) conducted a large-scale evaluation of over 200 academic physics programmes at each of the 13 Dutch universities as part of the Dutch national research quality assessment procedure. Each programme was judged by an international review committee that provided the basis for strategic decision making in research policy management (Rinia et al., 2001; Westerheijden, 1997).

In parallel, members of the Foundation for Fundamental Research on Matter (FOM) and CWTS at Leiden University conducted a complementary bibliometric analysis and compared the results with the VSNU expert committees (Rinia et al., 2001). The bibliometric analyses confirmed the results of the quantitative peer assessment to a statistically significant degree (van Raan, 1996). The bibliometric study also demonstrated there was no bias with respect to interdisciplinarity, i.e. peer judgement of physics papers published in non-physics journals show the same level of agreement with bibliometric indicators as physics papers published in journals within the physics field. Interestingly, for papers

published by authors in the condensed matter physics programme, regression analysis showed the strongest consensus between peer judgement and article-level citation indicators and only weak agreement between expert opinion and journal-only indicators (Rinia et al., 1998). This was considered important evidence to support the role of bibliometric analysis in research evaluation. Meanwhile, the latter finding supported the notion that journal impact factors alone should not be relied upon to predict the quality of individual papers published in selected journals.

Soon after these pioneering bibliometric studies, a major change in global society occurred that would have far-reaching consequences in the development of bibliometrics – the World Wide Web was invented. The World Wide Web was designed to create a pool of human knowledge (Berners-Lee et al., 1994) and it immediately became clear that it offered huge potential to the scholarly community, especially in navigating forward and backward in time through articles related to each other by cited references.

1.4.2 Impact of the World Wide Web

It was sometimes said at ISI that Garfield's citation index waited 40 years for the World Wide Web to be invented. Indeed, soon after the advent of the World Wide Web, the Web of Science was launched, which comprised the Science Citation Index joined by the Social Sciences Citation Index, the Arts & Humanities Citation Index, and later others. Thousands of universities around the world paid subscription licences so their academics could navigate the citation links and discover papers relevant to their work. Journal editors and publishers recognised the value of being on the inside and still today routinely submit journals for scrutinization against the Web of Science selection criteria. Only those journals that pass the test will be indexed, while others are told they may re-apply in a couple of years. Once a journal is indexed in Web of Science (or in other similar citation indexes, in particular Scopus), its content becomes more easily discoverable to academics who have access to the index. Some publishers see value in this increased visibility and submit their journals to the various selection and inclusion processes set out by the owners of Web of Science and other similar citation indexes. As university ranking systems often use citation indexes as the basis for the bibliometric component of their ranking calculations, some universities incentivise their academics to publish in 'indexed' journals. From the beginning of citation indexing, it was known that journals in different subject fields exhibit different publication and citation behaviour and therefore need to be treated according to the norms in their field (Garfield, 1983). I will now discuss the classification of journals into subject areas.

1.4.3 Science classification systems

Science classification systems developed that revolved around journals as nodes in the system connected by citations between them (Small, 1993; Small et al., 1985; Small & Sweeney, 1985). In the Web of Science, the journals are classified into one or more of approximately 250 subject areas and Scopus has developed a similar system. In both cases, a scholarly paper is not directly assigned to a subject field, rather all the articles in any given journal are assigned to the subject field that the journal has been assigned to. That means the system is not very sensitive because journals are not always strict about the precise field they cover, and many papers address topics that cross boundaries between subjects or are multidisciplinary in their nature.

The inherent differences in publication and citation behaviour between subject fields means that the performance of academics in one field cannot be directly compared with those in another field using simple bibliometric indicators. The top listed journals in the medical sciences have higher Impact Factors than the top journals in the social sciences. Newer journal indicators were developed that take into consideration the context of the subject field in which researchers publish their papers. An example of such an indicator is the Source Normalized Impact per Paper (SNIP) (Moed, 2010).

Alternative classification systems have been developed whereby articles are attributed directly to subject fields based on their contents. Dimensions uses the article-level classification system which cuts out the need to use the journal as the unit of aggregation. In a comparison between journal-level subject classification in Web of Science and the article-level system in Dimensions the authors claimed that the journal-level system in Web of Science was more accurate (Singh et al., 2020).

1.4.4 Bibliometric datasources

In every bibliometric analysis, the data to be analysed has to be defined and collected. Initial studies used journals indexed in the Science Citation Index but following the launch of several other data sources bibliometricians have a choice. The choice of data source is important and can influence the results of any study. Indeed, comparison of bibliometric data sources is a frequently discussed topic within the field of bibliometrics. It is useful therefore to describe the available data sources and I shall use this section to introduce the main ones.

The Web of Science is made up of component databases. The following databases make up the core collection of Web of Science: Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index (Science and Social Science & Humanities editions), Book Citation Index (Science and Social Science & Humanities editions), and the Emerging Sources Citation Index. The databases in the core collection are usually used when conducting bibliometric analyses based on the Web of Science. However, there are also regional databases including the Arabic Citation Index and Chinese Science Citation Database, subject-specific indexes such as the BIOSIS Citation Index, and indexes of non-journal literature such as the Data Citation Index.

In 2004, the publisher Elsevier launched a rival bibliometric database, Scopus, by combining several existing databases including Medline, Compendex, and Embase. This created a competitive market for access to global, multidisciplinary discovery services and bibliometric data that continues to this day. Web of Science and Scopus have each developed a coverage policy that determines which journals, conferences, and books they include in their databases. Web of Science comprises publications from high impact journals and conferences selected by an internal editorial team and focuses on quality (Birkle et al., 2020). Scopus is more inclusive, comprising publications from over 39,000 serial titles and relies on an external content selection and advisory board (Baas et al., 2020). Although there is considerable overlapping content between Web of Science and Scopus, important distinctions remain due to differences in coverage policy, indexing practices, subject field definition, publication date discrepancies, and differences in author name disambiguation, affiliation unification, and document type assignation.

Also in 2004, Google launched its own citation resource known as Google Scholar with almost universal coverage, no date limitation, and the huge advantage that it was freely available. That opened up access

to millions of academics all over the world, especially those not working or studying at a university with a Web of Science or Scopus subscription. Google Scholar has become widely used as an Internet discovery tool, but not as a data source for bibliometric studies, mainly because the underlying data is not made available. Google prevents automatic retrieval of its data using technology such as captcha that can only be solved by humans (Else, 2018). Another consequence of Google Scholar's inclusivity is that it may not be sensitive to the differences in output quality as determined by selective databases (Diem & Wolter, 2013) and therefore not an ideal source for evaluation.

In this context, Digital Science launched Dimensions in 2018. Dimensions seeks to harness the best of both worlds (Hook et al., 2018). Largely based on Crossref, Dimensions is more inclusive than Web of Science and Scopus, offers a limited free version, and extends the bibliometric database to cover a greater portion of the research lifecycle. Dimensions comprises data on research inputs such as grants, along with outputs including datasets, publications, and patents, also links outcomes such as social media impact, citations, policy documents, and clinical trials (Herzog et al., 2020).

Crossref is an organisation that generates digital object identifiers (DOIs) and assigns them to academic publications. Scholarly publishers can become Crossref members, which entitles them to upload article metadata from their journals into the Crossref database, whereupon each article is registered to a unique DOI (Collins, 2022). Members are encouraged to link the articles' URL and other metadata with the DOI and update the URL links whenever the article is moved. Regardless of where the document is posted on the Internet, the DOI remains the permanent and unique identifier for the document. Currently, Crossref comprises metadata for more than 100 million records and is adding metadata over 10 million records per year. This metadata is openly available, which makes it an attractive source for bibliometric study (Hendricks et al., 2020).

Following some high-profile cases in which journal selection policy was implicated in producing suspicious university ranking results, Microsoft launched Microsoft Academic Services (MAS) as a fully open scientific knowledge graph (Sinha et al., 2015). MAS collects data from the entire public web and includes studies at all stages of publication to avoid sampling bias and the associated potential impact on bibliometric studies (Wang et al., 2020). MAS does not use DOIs or ORCIDs because of the view that such technical standards are not well used by the scientific community and fail to live up to their promised consistency (Wang et al., 2020). A large-scale comparison between major databases showed MAS had far broader coverage than Web of Science, Scopus, Crossref, and Dimensions (Visser et al., 2021; Wang et al., 2020).

Microsoft Academic and related services were discontinued in January 2022 and OpenAlex was launched to fill the void. OpenAlex is named after the world's first library in Alexandria, Egypt and aims to improve transparency of research evaluation (Piwowar et al., 2022; Priem et al., 2022) by linking scholarly works, authors, venues, institutions, and concepts. The main data sources are the discontinued Microsoft Academic and Crossref along with preprint and data repositories such as arXiv and Zenodo. OpenAlex currently comprises 243 million works (OpenAlex, 2023). While some of the current data sources remain entirely or partially behind a paywall, OpenAlex is a fully open data source with open API and open source code (Priem et al., 2022).

There are many other bibliometric datasources available, and this section cannot mention them all. Some newer global, multidisciplinary sources have not been included in this overview and there are a multitude of region-specific and discipline-specific data sources available. Those mentioned above

should be taken as prominent examples of bibliometric data sources in a growing field, rather than an exhaustive review.

1.4.5 Applied bibliometrics – University rankings

Journal metrics such as the Journal Impact Factor have been used to rank journals according to their citation impact leading to competition among publishers to maximise the performance of their journals. In the last 20 years, bibliometric indicators have been used to rank the performance of universities in published league tables or university rankings.

In 2003, the Academic Ranking of World Universities launched the Shanghai Ranking, which quickly stimulated lively debate on the methodology used and its validity (Liu & Cheng, 2005; Van Raan, 2005). A paper entitled *Should you believe in the Shanghai Ranking?* comprises a scathing review of irrelevant criteria, a methodology 'plagued by major problems', and a poorly structured system (Billaut et al., 2010).

The following year, Times Higher Education launched the World University Rankings in cooperation with QS, which eventually split into two competing ranking systems. Each of these ranking systems has developed its own methodology based to varying degrees on bibliometric indicators alongside university-provided metrics such as faculty-student ratio, and surveys that quantify the opinions of the scientific community on the prestige of top institutions. More ranking systems were launched, and fierce competition ensued between bibliometric data providers to become the trusted database partner of the most popular rankings.

Controversy followed (e.g., Gadd, 2020) and moderate critics of university rankings proposed an alternative system in which universities are listed with a range of bibliometric and other indicators and the rank can be changed by the user depending on the criteria they select. A prime example of this is the Leiden Ranking (Centre for Science and Technology Studies, 2022; Moed, 2017; Waltman et al., 2012). In the absence of an overall or definitive ranking, this system can also be known as university profiling.

1.4.6 Community-led initiatives to improve use of bibliometric techniques

The argument about interpretation and use of the Journal Impact Factor raged for many years, until in 2013, following a meeting of the American Society for Cell Biology, 150 scientists and 75 organisations signed a statement calling for an end to misusing metrics to evaluate impact of individual articles and the performance of those who wrote them. At the time of writing this dissertation, the declaration has been translated into 30 languages and signatories number more than 20,000 individuals and 2,850 organisations (*DORA*, 2023). Among the signatories it is likely that a substantial proportion have signed to create the image that they care about responsible use of metrics but continue using the JIF just as before (Torres-Salinas et al., 2023). However, there was no organised resistance to the DORA declaration and no widely communicated counterargument. Even the company that publishes the JIF, Thomson Reuters (now Clarivate) was relieved that the broader scientific community was finally questioning misuse of quantitative indicators and was willing to engage in discussions on newer, more customised impact indicators.

As access to bibliometric databases such as Web of Science and Scopus increased and the owners of these databases launched their corresponding analytical tools, InCites and SciVal, organisations all over the world increased their reliance on bibliometric indicators. The university ranking systems, partially based on publication and citation metrics, gained further popularity to the point where in some parts of the world, ranking methodologies have influenced, or even de facto replaced, university strategy. In 2015, the Leiden Manifesto (Hicks et al., 2015) set out ten principles that describe best practices for metrics-based research assessment.

Similarly, in July 2022, the Coalition for Advancing Research Assessment (CoARA), started by the European University Association representatives along with Science Europe and the European Commission, published the agreement on reforming research assessment (ARRA). Widespread concerns about assessment based on oversimplified metrics have incentivised the European research community to call for an agreement between those being assessed and those conducting the assessment to lay some ground rules (CoARA, 2022b). The agreement envisions assessment of research, researchers and research organisations which is necessarily performed through qualitative peer assessment and supported by responsible use of quantitative indicators.

1.4.7 Network visualisations

The enormous growth in sheer numbers of journals, articles, and academics has created challenges for people tasked with evaluating science and scientists. Faced with large numbers of documents and people, it is sometimes easier to detect links between them when publication networks are presented visually as maps. One bibliometric mapping tool, VOSviewer, is specially designed to enable users to create maps based on network data and to visualise and explore the maps (Waltman & van Eck, 2012). The user can choose which attributes are presented on the map, for example author names, affiliations, or keywords. Chapter 3 of this dissertation compares methodological approaches to identifying research related to the UN sustainable development goals (SDGs) and uses VOSviewer maps to illustrate the difference between clusters as related to climate change using keywords from the title and abstract of the papers.

The opportunities of the Internet combined with the rapidly growing quantity of scholarly articles made network visualisation attractive to bibliometricians. Every research paper can be a node on a map and citations can serve as the links between them. Alternatively, papers can be represented by the keywords mentioned in the title, abstract, or full text and term maps present words more prominently the more frequently they are mentioned. As an example, in Figure 1.1, the most frequently occurring terms in the five articles in this dissertation are presented using VOSviewer. The terms fall roughly into five clusters related to topics studied in the articles and denoted by the different colours. The terms in purple, such as *country*, *country affiliation*, *country name*, and *population,* are all related to geodiversity, while *conference*, *asean country*, and *Indonesian paper* land in the blue cluster because they are related to the work on policy and behaviour. Some of the terms appear to fall between clusters, such as *SDGs*, *action*, and *climate,* which are somewhere between the red and green clusters. The terms indicate a close relationship between the topics of stakeholder collaboration and SDG classification method. Terms at the centre of clusters, such as *affiliation discrepancy* (data quality) and *triple helix* (stakeholder collaboration), are specific to their cluster and are only weakly related to the other topics.

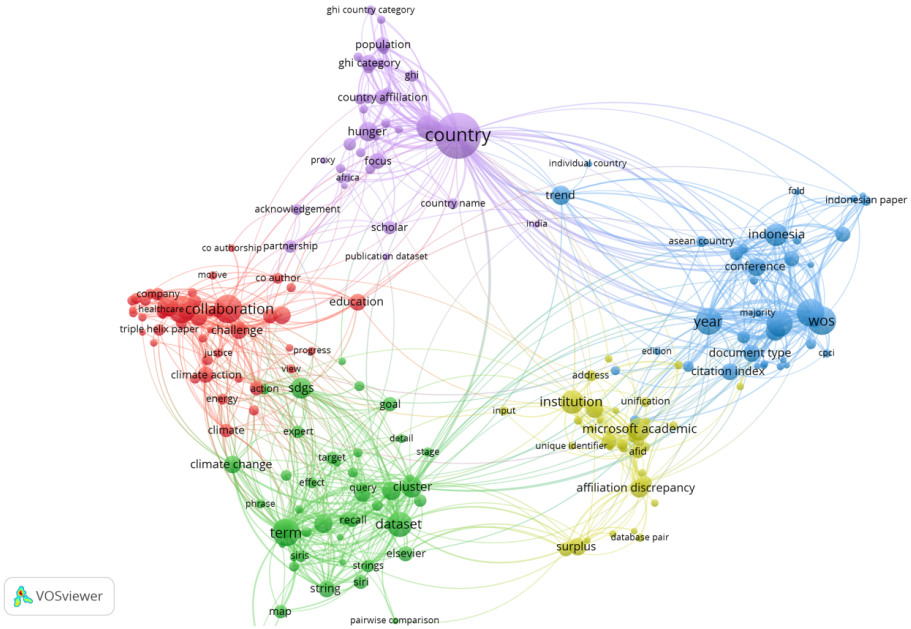

Figure 1.1 VOSviewer term map using the full text of the five chapters of this dissertation. Click here for live navigation

It is easy to see the relationships between concepts described above because we are looking at a map. The analysis of large numbers of terms is simplified through visualisation.

1.5 Motivation and objectives of this dissertation

1.5.1 Motivation

There is a need for clearer guidance on how bibliometrics can be used to support research policymaking and research evaluation. Parts of the scientific community support the use of quantitative indicators in research policymaking and evaluation. Supporters of the bibliometric approach cite its usefulness in relatively quickly and cheaply analysing large amounts of data and reliance on documented methods and indicators. Critics of the quantitative approach point to problems with bibliometric data accuracy and consistency and claim that only fellow academics are qualified to judge research quality. The drawbacks of both peer review and bibliometric approaches require improvement to gain support for use in policymaking and research evaluation.

In reality, neither approach on its own is likely to be ideal for most scenarios and the third option is to use a combination of quantitative bibliometrics and qualitative expert judgement. However, bibliometric data, methods, and approaches are not yet reliable enough to win the overwhelming trust of the broader scientific community. Data quality and curation need to be improved, and the community needs to play a greater role in the development of new methodological approaches and technologies. This dissertation addresses these points through five bibliometric studies that show both the power and the limitations of the quantitative approach. Areas for improvement are identified and solutions

proposed that will improve the reliability, trust, involve the community, and provide better guidance to stakeholders on the use of bibliometrics in the context of research policy and research evaluation.

1.5.2 Objective

I have picked five areas that show different ways in which bibliometrics can support research policymaking and research evaluation. The objective is to illustrate a range of scenarios where large-scale, quantitative bibliometric studies offer benefits to evaluators and policymakers. The studies also highlight the limits to bibliometrics and alert stakeholders to potential problems caused by misusing bibliometrics. The overall ambition of this dissertation is to present a series of recommendations for research policymakers on how to employ and interpret bibliometric techniques, and to provide suggestions to the wider scientific community on how bibliometric data, processes, and methods can be improved.

1.6 Structure and research questions

There are many areas in which quantitative research evaluation and research policy interact. The next part of this dissertation comprises five studies that address the objectives described in section 1.5.2. Chapters 2 and 3 present studies that illustrate limitations in data quality in bibliometric databases and subject classification techniques and cautions policymakers against over-reliance on the results of university analyses or rankings. Chapter 4 comprises a large-scale bibliometric study on the academic sector's transdisciplinary collaboration with societal stakeholders. Chapters 5 describes the geodiversity of research and discusses the extent to which bibliometrics can uncover unfair collaboration practices in developing regions of the world. Finally, chapter 6 illustrates a clear case of bibliometric evaluation causing a questionable change in behaviour of a national scientific community. The chapters address the following research questions:

1.6.1 RQ1: How does bibliometric data quality impact research policy?

Chapter 2 addresses the research question on data quality by illustrating wide-ranging discrepancies between four major bibliometric databases in their unification of author affiliations, an important type of bibliometric data. Author affiliations are used to link publications to universities or other research institutions, and it is therefore crucial to unify as many variations of the organisation name as possible. Each of the databases in this comparison uses a different unification system, which means they each produce a different publication set for the same university.

The findings draw attention to the limits in reliability of any university evaluation or ranking system. Understanding of the limitations of bibliometric databases and the discrepancies between them should be a prerequisite for policymakers before they design bibliometric studies of universities or act on their results. This study therefore presents author affiliation discrepancies between databases as an example of how data quality can influence the outcome of policy goals. The chapter concludes by expressing support for a single, community-led affiliation unification initiative such as the Research Organization Registry (ROR). ROR uses a unique identifier for universities in a similar way to the DOI for publications and ORCID for authors. Universal adoption of such a system for universities would channel community efforts into improving data quality and reduce discrepancies between databases.

1.6.2 RQ2: How can bibliometrics help improve the categorisation of sustainability research?

The research question on categorising new research fields is answered in chapter 3 through a comparison of four approaches (Elsevier, Digital Science, STRINGS, and SIRIS) to classifying research publications related to one of the sustainable development goals (SDGs) defined by the United Nations: SDG 13 (climate action). The analysis explores the different steps used by each method and suggests explanations behind the low agreement between the methods.

Sustainability policy has filtered down from the UN to government and university level and pressure is building for entities to report on their contribution to progress against the SDGs. Academic and research institutions have a need to produce bibliometric reports on their publications that show productivity, impact, and collaboration. Sustainability had not been delineated as an academic subject until recently and all the major bibliometric database owners, as well as academic groups, are racing to develop new methods of identifying research papers related to each individual goal.

The SDGs are multidisciplinary, complex, and vague, which means there is no precedent, no pre-existing categories, and no 'ground truth'. Each of these approaches produces different pictures of what an SDG publication dataset should look like, which leaves policymakers wondering which one is right. Chapter 3 contributes to the debate among policymakers on the extent of progress against the SDGs.

1.6.3 RQ3: What can bibliometric analysis tell us about transdisciplinary research collaboration trends?

The research question on transdisciplinary research is addressed in chapter 4 by using bibliometric data to define research output for the academic sector and three major societal stakeholders: government, industry, and nonprofit organisations. The study collects papers published by the academic sector over a 10-year period and presents the changing share of collaborative papers with each of the three societal stakeholders. Given the diverse nature of these actors in different regions of the world, the study then presents the results for the countries with the largest scientific output. This enables us to discuss the findings in the context of transdisciplinary research at national level.

It is noted that different stakeholders often have different motives for collaboration, which could actually inhibit joint research projects. The goal of academic researchers is almost always to publish their results to share knowledge, foment discussion with fellow scholars, contribute to solving problems, and as a means to advancing their careers. Industry on the other hand, is ultimately driven by bringing new products and services to market, which sometimes requires research findings to be kept secret and other times only to make their findings publicly available by patenting their inventions, rather than publishing them in academic journals. These conflicting goals may act as inhibitors to large-scale academia – industry collaboration. Differences between academia and government research focus may produce similar conflicts.

The relationship between stakeholders may therefore be lopsided and limit the amount of collaborative research conducted. Therefore, policies that simply incentivise 'more inter-stakeholder collaboration' might be too simplistic. Research policymakers around the world are calling for closer collaboration and highlighting the perceived benefits of transdisciplinary research, but data on progression is scarce.

Chapter 4 therefore contributes to the current literature on transdisciplinary research by providing feedback to policymakers on the success of their efforts.

1.6.4 RQ4: How can bibliometric techniques be used to describe the geodiversity of research?

The study presented in chapter 5 offers insights into the geodiversity of research through bibliometric analysis of the relationship between research focus and author locations. The study uses SDG 2 (zero hunger) as a case study and uncovers wide variation between geographical topic focus and author location depending on the level of hunger in the country of focus.

Chapter 5 discusses policies emerging among publishers and academic societies that seek to ensure fair and equal partnership between local academics in countries severely affected by hunger and visiting researchers from wealthy countries. Such policies deal with local capacity building, treatment of samples, and the order of co-authors on the resulting publications. This is a case of the publishing community setting policies that could influence academic behaviour and be visible through bibliometric reporting.

Bibliometric studies can help identify the existence of problems through illustration of geographical topic focus and author affiliations using mapping visualisations. Country representation and author position can contribute to our understanding of collaboration dynamics and time series will enable communities to follow progress through trends. Through bibliometric studies, policy makers can identify the existence and extent of problems and monitor progression once steps have been taken to address them.

1.6.5 RQ5: How can bibliometric assessment cause inadvertent behavioural change in the scientific community?

The research question on the effects of bibliometric evaluation on academic behaviour is addressed in chapter 6 through a bibliometric study on published document types in 10 countries in Southeast Asia over a 20-year period. The findings revealed a national change in publishing behaviour specific to Indonesian academics and linked it to the introduction of a specific research policy. The Indonesian government introduced incentives into its national research policy with the intention of stimulating a general increase in academic publishing. However, the academic community responded to the policy by switching from publishing articles in journals to publishing papers in conference proceedings. Conference papers were perceived to be quicker and easier to publish than journal articles and therefore conference publishing was perceived to be a preferable route towards achieving the policy incentive.

This chapter highlights the potential consequences of using a metric as an incentive. The Indonesian study is a classic case of Goodhart's Law, which states that when a metric becomes a target, it ceases to be a useful metric (Goodhart, 1975). The chapter draws parallels between the Indonesian case and the international university ranking systems. Universities in many parts of the world see a rise in the rankings as a mark of success. As academic publications and citations contribute to higher ranking, bibliometric indicators become targets.

Chapter 6 discusses the implication of these findings in that policymakers overlooked the difference between the characteristics of journal articles and conference papers when developing the research policies. The scientific community realised this and accelerated submissions of papers to both national and international conferences whose proceedings would be indexed in bibliometric databases. This is an example of policymakers using bibliometric data to stimulate one change, but inadvertently causing another.