



Universiteit
Leiden
The Netherlands

Bibliometric uncertainty: data quality, reliability, and validity

Donner, P.

Citation

Donner, P. (2026, October 1). *Bibliometric uncertainty: data quality, reliability, and validity*. Retrieved from <https://hdl.handle.net/1887/4310427>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310427>

Note: To cite this publication please use the final published version (if applicable).

1 Introduction

Any large-scale research assessment exercise, as well as any mapping of research fields for sociological or science policy purposes, is heavily conditioned by the quality of the information retrieval and data-cleansing routines applied to the source database. (De Bellis, 2009, p. xxii)

This dissertation is a methodological contribution to bibliometric research building on empirical investigations. In the four original research chapters, I critically examine important aspects of the fundamental validity of the conventional approach to core bibliometric methodology – publication and citation analysis. I introduce newly developed methods to overcome existing shortcomings and to improve the reliability and validity of citation analysis, and to express bibliometric indicator uncertainties explicitly that follow from imperfect data quality.

To get to the question of how to better handle data errors and inaccuracies in bibliometrics, I will begin this dissertation with the more fundamental question of why bibliometrics and metrics-based research assessment are done at all and why they are done in the particular way they are nowadays done. To this end, I need to sketch the sociopolitical environment in which bibliometrics developed and which it continues to be part of.

1.1 The socio-political environment of bibliometrics

Scientific research has over the course of the twentieth century become such an integral part of modern societies that its current organization, in which most basic research is carried out by fully or largely publicly supported professionals in research universities and dedicated research organizations, might be taken for granted. In financing academic basic research on a grand scale, the financier, usually a state, would do well to find some means to assure that its subsidies are not squandered or used too inefficiently. If that were to happen, public opinion may turn against, not only the subsidized research system, but also against that part of the state that has allowed things to come so far and which the public assumes to bear responsibility for the diligent and efficient use of state funds. All state funds are, after all, resources generated by the public through their economic activity but requisitioned by the state through taxation

and thereby withheld from private consumption and saving and directed towards consumption as directed by the state. A state thus finds itself in a position where it must exercise oversight over the spending of its dispensations to academic research lest it risks drawing the ire of the public on itself for neglecting to do so if things go awry (Elzinga, 2012).

In pursuit of this objective, states face resistance by the organized profession of researchers which insists on a large degree of autonomy in what it considers matters internal to the profession, such as personnel decision-making, priority setting of research agendas, selection of projects for funding and demarcation of science from non-science. The profession rejects outside steering as undue interference on the grounds that such processes require a degree of specialized knowledge and tacit experience that non-professionals do not possess, a claim that Bimber & Guston (1995) (p. 557) refer to as the Platonic exceptionalism of science. Although existing for much longer, over the course of the past decades, the family of judgement and decision-making procedures which can be subsumed under the label of formal peer review has attained a central position in practical but also in ideological terms due to its embodiment of the terms of the discourse as set out by the research profession: scientific and scholarly judgement can only be legitimate if exercised by qualified peer researchers. Outsiders, such as politicians and bureaucrats, are held at arm's length by the profession by arguing that if science is to achieve its goals, including objectives commissioned by outside societal consensus, the most rational way to do this is by relying on specialist scientists' judgments, by peer review.

In contrast, applied bibliometrics, has partially taken on the function of a technology and a means of state legitimization¹. It has that function, even if it perhaps was not explicitly and consciously developed for it and if perhaps some practitioners find this characterisation distasteful. It is hence neither surprising that bibliometrics in its incarnation as a basic science hardly studies non-academic, industrial research, nor that bibliometrics in its application in research evaluation primarily finds use in the assessment of publicly funded research rather than privately funded research, be it private enterprise or private philanthropy, precisely because these types of research financiers are not

¹Godin (2005) put it this way: "But most often, governments constructed statistics to serve *symbolic, ideological* and *political* ends. Statistics certify the user's credibility and seriousness. It objectifies political choices, that is, eliminates arbitrariness and inherent subjectivity in political action by portraying decisions as transparent and lending legitimacy to political discourse. S&T statistics is no exception." (p. 5, italics in original)

accountable to the public. It is not easy to imagine what bibliometrics would look like in the absence of large-scale state funding in research-based knowledge production, but it is obvious that it would appear very different. This does not mean that bibliometrics has no *raison d'être* independent of state demand for research evaluation and oversight, only that this demand has historically shaped its development (Godin, 2005; Wouters, 1999, p. 131).

One can find acknowledgment of this connection between knowledge production and the use of bibliometrics for research evaluation in the professional bibliometric literature. For instance, Vinkler (2010) stated that “The aim of scientometrics is to reveal characteristics of scientometric phenomena and processes in scientific research for more efficient management of science” (p. 1). De Bellis (2009) wrote:

From a science policy standpoint, it [the recognition of scientific quality] is also a useful guide in the allocation of the financial and human resources necessary for the machinery of science to work properly. (p. 181)

The peer review crisis triggered off the search for alternative or complementary evaluation methods more detached from subjective judgments. So, at least from the 1970s, the use of citation analysis to produce indicators of scientific performance in terms of citation scores began to receive unprecedented attention from politicians and science managers, clearly attracted by the several virtues of a numerical index: convenient, quickly understood, easily applied, and, above all, easy to calculate thanks to ISI databases. (p. 183)

An OECD working paper on bibliometric indicators stated:

A change took place in the 1970s: science was no longer seen as a venture in which society could invest generously and without limit. [...] science and technology were expensive, but investment in research did not automatically make it possible to solve environmental problems or social problems [...] This led to concern over the profitability of basic research in particular, and researchers were increasingly perceived as “producers of science” who had to account for the funds

they received. [...] The new aim was to produce value added that conserved natural resources and created less pollution, and to create a more efficient research system that made better use of existing intelligence. It was in this context that the evaluation of scientific research came into its own. [...] methods had to be found to quantify, compile and compare indicators. The establishment of a measurement of science(s) became inevitable. This shift cleared the way for the analysis of science and technology and favoured the advent of bibliometricians in science policy. (Okubo, 1997, pp. 11–12)

The emergence of bibliometric indicators – as primarily a science policy instrument – is traced historically by Godin (2005), chapter 6, to the publication of the first *Science Indicators* report of the US National Science Foundation in 1973, developed by its National Science Board. It was widely read and discussed and sparked the publication of various national and international report series that followed its model. The first report included publication and citation statistics calculated from the Science Citation Indexes which were prepared by Computer Horizons, Inc.

While bibliometric indicators are presently used extensively also within the semi-autonomous research system (see section 1.5.3) they were up to a point historically primarily developed as a tool for monitoring developments in the research systems from outside. Bibliometrics could serve as a tool to counterbalance a progressing emancipation of a societal subsystem. Complete independence and unchecked self-administration of science is not in the interest of states, which would continue to bear the responsibility of financing increasing demands while losing the wherewithal to verify the prudent spending of funds. The logic of rationality that bibliometrics derives from (purported) objectivity and quantification makes it an ideal tool of observation, steering, and decision-making at arm's length, from outside the system itself. A further advantageous feature of bibliometrics as such a tool is that it works, at a first approximation, without directly interfering with the day-to-day conduct of scientific research, that is, it does not infringe on the prerogatives of the profession.

We shall take the large-scale, state-mediated public funding of scientific research and the role that bibliometrics has assumed in this environment as given for this dissertation². We shall ask, *given* financial state support for academic research,

²Suffice it say that economists are by no means in unanimous agreement that government

given the need for accountability consequently demanded by the public bearing the costs of these expenses, how can bibliometrics contribute meaningfully to such oversight and how does the quality of bibliometric data matter for such a purpose.

1.2 Terminology

Bibliometrics is one of several methodological approaches to science studies or scientometrics, that is, the scientific investigation of the socio-cognitive process of scientific activity itself.

The term ‘bibliometrics’ was popularized by Alan Pritchard in 1969 (Pritchard, 1969) although its French version, *bibliométrie*, was used earlier by Paul Otlet (Rousseau et al., 2018). Pritchard defined bibliometrics as “the application of mathematics and statistical methods to books and other media of communication.” Historical overviews of bibliometrics, corroborating the close co-development with large-scale involvement of states in supporting basic research can be found in Van Raan (2019) and De Bellis (2009). Van Raan (2019) defines bibliometrics this way: “Within scientometrics, the research on scientific communication, in particular with data from publications, citations, and journals, is called bibliometrics”, with scientometrics being the “quantitative study of science”. A comprehensive discussion of the terminological questions of this branch of science can be found in chapter 1 of De Bellis (2009).

Bibliometrics as a complex of methods and tools lends itself both to the curiosity driven examination of the process of scientific research itself, by taking its communication system as a source of data, and to applied use. Applied use of bibliometrics can range from purely documentary and informative statistics to evaluative assessment with important science-political consequences, such as funding decisions.

As a largely empirically operating discipline, bibliometrics requires data of sufficient data quality in order to produce meaningful and useful knowledge. While this may seem like a truism and while the same holds true for any other data-driven discipline, data quality and its ramifications are nevertheless not routinely explicitly addressed in the substantive bibliometric literature despite a wealth of methodologically oriented studies showing severe and wide-ranging

intervention in science funding is a good idea (Kealey & Micheals, 2019).

data quality deficits of principal bibliometric data sources. Gauch (2021) goes so far as to state that “most bibliometricians are strongly inclined towards retaining a high level of data quality in the bibliographic data they use. Sometimes to an extent that may appear pedantic to the average user of these data sources” (p. 163). Perhaps, but analyses of the sensitivity of the results of studies to errors in the data are the exception in the field. Because not much is known about how much fundamental quality attributes of bibliometric data affect bibliometric indicator values, it is not possible to say whether data quality is generally good enough to make studies reliable and trustworthy or so poor as to call many results into question. Be it one way or the other, as long as this knowledge deficit exists, bibliometrics as a scientific discipline is wide open to criticism for relying on data of poor or even unknown quality, which in itself presents a vulnerability to its acceptance. In short, the overarching postulate of this dissertation is that the credibility of individual bibliometric studies, of bibliometrics as a research endeavor, and of bibliometrics-based research evaluation rests inextricably upon sufficient data quality.

The research contributions presented in the substantive chapters of this dissertation demonstrate empirically that bibliometric data quality of commonly used data sources with commonly applied methodology is often poor. So poor as to question present bibliometric practice, which is largely indifferent to appreciation of uncertainty in results caused by data quality. The contributions of the dissertation go beyond this in proposing and demonstrating methods to improve data quality and a method to propagate and express information on uncertainty of bibliometric indicators that arises from data quality. The remainder of this introductory section will provide the necessary context for the substantive studies and the motivation for studying bibliometric data quality.

1.3 Conceptions of data quality

1.3.1 Definitions

Parts of this dissertation constitute applied data quality research. That is, empirical research on the quality attributes of specific data and on the consequences of data errors for the use of this data. What, precisely, is meant by the term ‘data quality’? We proceed to discuss the concept of data quality and some dimensions of data quality that are considered of central importance in

the literature and that are also important in bibliometrics.

The international norm ISO 8000-2:2022 defines ‘data quality’ as the “degree to which a set of characteristics of data fulfills requirements” (3.8.1). ‘Data’ is defined (3.2.2) in the same source as “reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing”, while a ‘requirement’ is a “need or expectation that is stated, generally implied or obligatory” (3.2.1).

1.3.2 Demands of data quality in basic versus applied bibliometrics

An understanding of the meaning of data quality in bibliometrics can build on this basis. Bibliometric data relate to the human activity of scientific research, in practice often with an emphasis on data relating to outcomes or end products of this activity – scholarly communication. Requirements for bibliometric-scientometric data can be derived from the objectives of bibliometrics. Bibliometrics is not one indivisible category of activity. It has developed into a wide-ranging and complex sphere of activities with many facets and aspects, competing schools, conflicting motivations, and diverging orientations. Like with other scientific disciplines, some parts of this complexity can be simplified by focusing on one of several dimensions by which bibliometric research may be ordered – the basic research-applied research dimension. In bibliometrics, both investigative basic research, which aims to further our understanding of the workings of the human social activity of scientific research, coexists with applied research. This is not to say that any bibliometric study must fit somewhere on a pole or a specific point in between the poles on this spectrum. Any given study may have aspects of both basic and applied research. However, frequently a clear positioning of studies on the basic-applied continuum is possible and may be useful with respect to considerations of data quality.

In the basic, fundamental, curiosity driven area of bibliometric investigation, the goal is the attainment of relevant knowledge that is rather general and generalizable, not tied very closely to a very specific context in time, place, and culture. Here, the sought after knowledge often is in the shape of broad law-like patterns that recur again and again, statistical regularities and relationships, assessed as to their fit to some theory or demanding a theoretical explanation in the first place. For this type of bibliometrics, a certain degree of data quality deficiency may be acceptable, if it can be assured that the results are robust

enough. If any observed patterns are volatile to minor data quality deficits, this might even arouse the suspicion that they are idiosyncratic artifacts that cannot be reproduced in other contexts. Nevertheless, a fundamental level of good data quality is indispensable for generating useful basic bibliometric research results. What that required level of data quality is, may depend on the local requirements of the specific study, for instance the type of research producing units which is analyzed, where a somewhat lower data quality may be acceptable for studies of entire countries, compared to studies of individual researchers.

In contrast, when it comes to applied bibliometrics, i.e. research evaluation, assuring very high data quality is of utmost importance, as bibliometric reporting is expected to provide a reliable basis for decision-making. It is the responsibility of the bibliometric evaluation expert not only to make that assurance but also to clearly communicate the level of data quality, including its affordances and limitations, to the clients or decision-makers (Vinkler, 2010, p. 4). The values of bibliometric quantities must in this context be as accurate (and transparent) as possible to support fair decisions in high-stakes situations like funding distribution or personnel hiring, promotion, or compensation. In summary, the data quality requirements in applied bibliometrics can usually be expected to be much higher than in basic bibliometric research.

1.3.3 Dimensions of data quality

Data quality is commonly seen not as a monolithic, but as a compound concept with an internal structure. Batini & Scannapieco (2006) describe data quality as a multifaceted concept comprised of several dimensions. The most important data quality dimensions considered in the literature are accuracy, completeness, currency, and consistency. The (semantic) *accuracy*³, or correctness, of a data value is understood as the closeness between a value as it exists in an information system and a value which correctly represents the real phenomenon that the data value is a surrogate for, i.e. its true value (Batini & Scannapieco, 2006). For example, in bibliometrics, if a bibliometric database gives the publication year of an article as 2010 but from the original publisher data it can be determined with certainty that it was published in 2009, then the accuracy of that value is affected. Unwarranted duplicates of records in a database are another

³In contrast, the syntactic accuracy of a value refers to whether a specific value in an information system is a member of the admissible domain of values for a variable in that system.

instance of accuracy problems commonly encountered in bibliometric data (e.g. Valderrama-Zurián et al., 2015). These may result from errors in data collection or processing. One example are errors in data disambiguation, such as in author name disambiguation, where one real person is inaccurately represented by more than one record.

The *completeness* of data refers to missing values that should be present because the entity or attribute exists in the real world and falls under the scope of the information system. Completeness is just as crucial for bibliometric data as for any other data. Consider for example the consequences in evaluative bibliometrics if values for attributes such as author name or author affiliation are missing in the information system which do exist on the original publications. The accuracy of values can only be determined for non-missing values, the accuracy of missing values is indeterminate. This data quality concept of completeness should be clearly distinguished from the more commonly used concept of data source completeness, or coverage. The latter refers not to a data quality dimension but to a property of a data source, namely to the proportion of a specific defined literature of interest that is included as primary indexed records in that data source. Bibliometric data sources usually have accessible registers of the covered literature. There is a situation when these two concepts intersect. If a data source claims coverage in its register of literature sources but does in fact not include all items of a source in the allegedly covered period, say all articles in a journal that is supposed to be covered, then this data source has a completeness insufficiency in both meanings of the concept.

Attributes of real-world objects change over time and so must the data representing them in order to be faithful to reality. Thus, *currency* is another relevant data quality dimension. Currency means how up-to-date the data is; whether it is current or outdated. Important examples for bibliometrics are updates to publication status such as retractions or the representation of newly introduced attributes for existing records. But also just the time it takes for newly published publications to be reflected by new records in bibliometric databases is an important data quality concern for users. A value that is no longer current is necessarily inaccurate, so currency may also be seen as a sub-category of accuracy.

Consistency refers to whether data in an information system or across information systems that are to be integrated are logically consistent with each other, that is, free of contradictions or conflicts. In many cases, requirements for data

consistency can be directly integrated as logical rules in a data system must that related data must fulfill to be considered technically consistent. Consider the fictitious example of a newly founded university, active as of 2025. If we find in a bibliometric database publications with affiliation data for that university from before 2025, this may indicate a breach of consistency.

Several more data quality dimensions have been proposed in the literature, cf. Batini & Scannapieco (2006). Some are slight variations on the major consensus dimensions treated above, some are domain-specific conceptions of data quality dimensions. Thus, the foregoing is not meant to be exhaustive but selective for the purposes of this dissertation. This dissertation is mostly concerned with the quality dimension of accuracy (in the strict sense), although completeness will also play a role in one chapter: Chapter 3 deals with publication month data, which is often not available in the used data but can be approximately inferred. A method to do this, and so to overcome the incompleteness, is proposed and evaluated.

1.3.4 Other distinctions of data quality in bibliometrics

Besides these internal dimensions of the concept of data quality, several other distinctions or categorization of data quality are relevant in bibliometrics.

Bibliometric data quality studies may examine the quality of scientific publication data at different *stages*, such as the original publications, their representations in the metadata available at publisher websites, the data on those publications in citation indexes or other information systems, or the derived data that bibliometricians and scientometricians create from those sources for their own specific purposes. In this dissertation, several of these stages will play a role. It will mostly examine data as it is presented in the raw data of citation indexes. However, I will also compare this to the more basic data from publisher systems and original publications in order to empirically assess levels of data quality. Moreover, also the further processed derived data that specialized bibliometric methods produce will be studied for its quality.

A very closely related, yet logically distinct way of thinking about bibliometric data quality is to differentiate according to the *source* of errors. An error, or inaccuracy, that is present in the original publication, say for example when authors cite a reference incorrectly, may be present at later stages of data use,

but it may also be corrected in later stages. The processing system of the Science Citation Index has had capabilities to detect such erroneous references and to correct them for a long time (Garfield, 1983). But such algorithmic manipulations may themselves introduce errors. This is how the distinction between stage of data and source of error becomes clear. The source of error may be a human or software. Different error sources may act at different data stages. The specific source of error will play a central role in Chapter 4, where my co-authors and I compare different affiliation disambiguation systems which in principle use the same basic data but generate outputs of different data quality.

Data quality takes on a higher level of complexity in data source integration, when multiple information systems are combined to supply a single end user access level. Daraio et al. (2022) propose an integration-layer approach to control data quality in data source integration, focusing on the data quality dimension of completeness, whose role in integration is highlighted. Their case studies from the realm of science policy studies data illustrate that data quality control at the integration layer relieves end users of considerable data quality analysis work by providing them with harmonized data that is by design as complete as possible given the completeness of the integrated systems.

Data errors in bibliometrics can also be classified into *random errors* and *systematic errors*. When errors are randomly distributed, they affect the possible precision of calculated indicator values. If data are affected by many random errors, any quantities calculated from them will have low reliability. Data of too poor reliability may be useless. Knowing the level of data quality –and hence reliability– is therefore essential. Systematic errors have some non-random component, and they cause specific deviations of observable calculated values from the values that could be obtained if all errors were corrected. As an example from bibliometrics one might think of the core processing step of reference matching. If a reference matching algorithm is relatively unsophisticated, it might systematically match some types of cited references worse than others. Author names are a crucial component of reference strings. Think of authors with complex names, such as compound names, names with accented letters, names from different writing systems which could be transliterated in different ways into Latin characters. Such complex names, and since names are culturally embedded, names from some cultures more than others, could lead to systematically more frequent failures of a reference matching algorithm to establish

a match where there should be one. In this hypothetical example, the citation counts for entire countries might be affected by greater systematic error, that is, they would be biased and show lower citation counts than warranted. This example was hypothetical, but the problem of systematic errors in citation matching is real. Franceschini et al. (2014) were able to identify systematically different error rates in citation matching for different publishers of citing papers.

In a sense, random errors are less severe than systematic errors, precisely because they are random and are therefore equally likely to affect any observation to the same degree and with the same probability. Systematic errors on the other hand affect observations with different probability and to different degrees. However, if the class of observations that are affected by systematic errors can be identified well and if the size of the systematic error is well-understood, systematic errors can be corrected for, while this is not possible for random errors. If, for example, citations for papers by authors with complex family names are systematically undercounted because of matching errors, this can be corrected for statistically by adding a factor to each citation count of this group of papers which corresponds to the size of the known undercounting.

This dissertation is concerned with both random and systematic errors. As an example, I study the accuracy of document type assignment in Chapter 2 and find that some document types are more likely to be incorrectly categorized than others, which is a form of systematic error. The accuracy of the citation counts is studied in Chapter 5. Imprecise or faulty referencing by authors, but also imperfect reference matching by citation index producers, can lead to unmatched references. On the level of individual publications, their degree of being affected by matching failure is probably largely random since a number of independent factors bear upon the end results. However, as the citation matching algorithms themselves are (presumably) deterministic, that is, they produce the exact same results for the exact same input, citation matching errors on the level of specific cited reference *strings* are arguably systematic – if a specific reference string is not matched or not matched correctly in one citing document, it will also not be matched in all other documents in which it occurs.

1.4 Data sources

Bibliometric studies usually draw upon rather specialized information systems, bibliometric-bibliographical databases of scientific publications. Their distinc-

tive characteristic is the inclusion of publication metadata of particular relevance for science studies, first and foremost citation data, which sets them apart from similar scientific literature database systems, such as research library catalogs. The citation indexing database, as a data source of bibliometric studies, will be the location at which data quality and bibliometric indicator uncertainty are studied in this dissertation. That is not to say that the data processed, recorded, and stored in these systems –metadata on the scientific publication system– is the only relevant type of data for bibliometric studies. Other types of data, especially data collected specifically for bibliometric research, such as a survey data, are also important. However, the scope of this dissertation is limited to bibliometric-bibliographical databases, or scientific citation index databases, which is justified by their historical and present prominence as primary data sources for bibliometric activity.

Data quality deficits relevant for bibliometrics can exist in the original publication before it is ever indexed in a bibliometric database. For instance, incorrectly cited references are a typical error category at this stage. But the bibliometric data source may also itself be the place in which errors in the data are introduced, when the originally present data are not correctly reproduced. This is because the processing of original data is not an error-free process. Empirical research has shown that bibliometric databases introduce numerous errors into the data, as will be discussed later on. This means that the choice of a bibliometric data source for a study has important consequences for the level of data quality and reliability of bibliometric indicators of that study.

Large scale bibliometric research and evaluation was for decades dependent on the Citation Index products of the Institute for Scientific Information (ISI), later Thomson Reuters, currently Clarivate Analytics. ISI's founder, Eugene Garfield, created the first multidisciplinary science information retrieval tool built around the idea of searching by cited references and citing documents, the Science Citation Index (SCI) (Garfield, 1955, 1964; Wouters, 1999). Citation Indexes were in time introduced for the social sciences, arts and humanities, proceedings literature and scientific books. These have been used since their early days for bibliometric studies of the science system and for decision-making.

Data quality was taken seriously at ISI from the very beginning of the SCI's history. This effort was led by Irving Sher, Director of Quality Control and Director of Research and Development (Garfield, 2001). Sher and colleagues reported on the implementation of a procedure for correcting references cited

incorrectly in papers in Sher et al. (1966). The following quote may give an interesting impression of what data quality work meant in the 1960s:

In the preparation of the IBM cards for the Science Citation Index, all keypunched data are verified by a second operator or reader. The IBM 557 interpreter with proof feature is used to interpret the cards and edit for erroneous multipunch characters and a preliminary field edit. After all the errors detected by the verifier have been corrected, an IBM 360 computer is used to edit and identify any further errors of the type which can be checked for by algorithm. This very elaborate computer procedure confirms card sequencing, field positioning, and acceptability of the types of data allowed in the various field positions. (Sher et al., 1966, p. 134)

Among the various relevant metadata fields, it is particularly the cited reference strings, with their great variability across disciplines and journals, that are challenging to process:

Normal citation behavior in hermeneutics, the interpretation of Scriptures, includes exegesis, the detailed documentation of sources. This is also expected of lexicographers, etymologists, and encyclopedists. Similarly, law review journals provide a level of detailed documentation that might be considered pedantic in other fields. [...] On the other hand, in engineering journals, pageless citation of textbooks, handbooks, and data compilations is frequent. In chemistry, the specific pages are cited but complete pagination for the full article will be missing. In citation indexes this results in the artificial separation of citations to the same article. These individually cited pages are listed as separate items, inflating the size of the indexes (Garfield, 1996, pp. 453–454)

The last few decades have witnessed massive changes in the citation index market. Commercial competitors such as Scopus (Baas et al., 2020) and Dimensions (Herzog et al., 2020) have been established. These databases are also used regularly as sources for bibliometric studies. The vendors offer, for subscription fees, raw data access to their products and access to bibliometric analytics platforms. Data from Google Scholar, a bibliometric data source that does not

require paid subscription, has also been used for studies, with noted caveats. Completely open citation data sets have recently been made available (Peroni & Shotton, 2020; Priem et al., 2022).

Several countries and regions have set up their own citation index databases to address their local needs. There exist or have existed the Russian Citation Index (Gorin et al., 2020), the Chinese Science Citation Database (Jin & Wang, 1999), the Chinese Social Sciences Citation Index (Hua, 2001), the Arabic Citation Index (El-Ouahi, 2023), KCI Korean Journal Database, and SciELO, covering Latin American research output – all hosted on Web of Science. SciELO is actually much more than just a regional Citation Index. While its citation index component is integrated with the Web of Knowledge since 2013, SciELO was founded in 1997 and it also exists as a freely accessible stand-alone platform. This platform also encompasses an entire open access publishing system. There is further the Serbian Citation Index (Šipka, 2005, also including a journal publishing system), Citation Index of Kazakh Publications (Suleimenov et al., 2009), the Indian Citation Index (Giri & Das, 2011), the Taiwan Citation Index - Humanities and Social Sciences⁴, the Citation Database for Japanese Papers (Negishi et al., 2004), the Islamic Citation Index (Mehrad & Arastoopoor, 2012), and MyCite (Malaysian Citation Index)⁵. As of 2022, the Armenian Citation Index was under construction (Sargsyan et al., 2022). Little is known about the data quality of these regional and national citation databases. Nearly all data quality studies have used the major international commercial databases⁶. Discipline specific citation indexes also exist, such as INSPIRE-HEP and Astrophysics Data System.

1.5 Indicator use and the importance of bibliometric data

In this section a broad overview substantiated by diverse examples is given to demonstrate the motivation for research into bibliometric data quality. It is argued that bibliometric data quality should be seriously studied because bibliometric performance indicators and bibliometric data have had a profound

⁴<https://tci.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=5gvP6n/aboutdb>

⁵I cannot confidently conclude that I have not overlooked further citation indexes.

⁶However, the data quality measures of SCIndeks - Serbian Citation Index for cited reference accuracy are notable. They include, in the integrated publishing system, automated checks of correspondence of in-text citations to the reference list and a system to alert authors to references to retracted papers, cf. <https://scindeks.ceon.rs/static/about.aspx?url=qualitycontrol.htm>

transformative influence on the way science is assessed and therefore also indirectly on how science is done and which science is done. Bibliometric evaluation under conditions of insufficient data quality can thus have considerable adverse consequences. Careless use of bibliometrics may lead to flawed or overconfident decisions made unwittingly on the basis of data that, when scrutinized, did not actually allow for clear-cut decisions or would have supported different ones. In other words, particularly in an evaluative context, the use of bibliometrics in a state of ignorance of underlying data quality may be an instance of irresponsible research assessment.

1.5.1 Non-opposition of qualitative and quantitative research assessment

The appreciation, understanding and communication of the value and relevance of scientific works and of individual researchers on the basis of their works can only be done to the general acceptance of those whose research is being judged by colleagues who are highly knowledgeable about the research field and whose conduct in evaluation is so transparently objective and fair-minded to be beyond reproach. Any detectable degree of partiality, be it favorable or antagonistic, is usually regarded as a breach of the norms of science as a social endeavor with the objective of the production of certified and reliable knowledge about the world.

The in-depth, qualitative evaluation of research outputs has limitations. When the necessity arises of assessing and comparing the research outputs of larger, productive research units and organizations, such as universities, or countries, formal peer review does not scale (Abramo, 2024, p. 4). However, as bibliometrics takes over when it comes to the assessment of thousands or hundreds of thousands of research publications, neither formal peer review nor its premises are left behind completely, rather it is supplemented by products of another type of appraisal by researchers.

It is apposite to reflect here on the sometimes overlooked connections between qualitative peer assessment and quantitative publication and citation analysis. Over the past decades, a social reform movement has been gaining ascendancy which is contesting the relative roles and balance of metrics-based and qualitative research evaluation and which has found expression in a number of successive declarations, manifestos, and initiatives (Rushforth & Hammarfelt, 2023;

Sivertsen & Rushforth, 2025). Within this movement, some have argued that qualitative peer assessment is generally preferable to bibliometrics in research evaluation, a proposition which I consider unwarranted given the identity of underlying cognitive processes.

What researchers do when they read scientific literature is always also a judgment process, an internal, informal, personal assessment. Researchers are, as Polanyi (1962) put it, engaging in a constant process of ‘coordination by mutual adjustment of independent initiatives’. This assessment proceeds by shared standards of judgment. The outcome of the assessment of published research by reading researchers may be a decision now or later to cite or not to cite the work that has been read or to recommend it to colleagues, depending on its perceived usefulness for their own research, which ties citations closely to the concept of research quality (Gläser, 2024). Such decisions will be made with some of the same motivations and by the same criteria as the assessment of works in formalized peer review settings, be it manuscript peer review before publication or post-publication peer review for purposes such as recruitment, promotion, institutional or national research output assessment exercises. The two different processes have in common the valuation of the methodology and findings of research according to their scientific soundness, originality, usefulness (Aksnes et al., 2019; Gläser, 2024; Polanyi, 1962) – but also interestingness, controversy, socio-cultural relevance, ideological alignment, etc. The difference between the settings is that the referee in the formal setting is making an assessment following such criteria as mentioned but addressed to and prepared for either a small group of decisionmakers or the wider scientific community, in a more or less standardized or at least normatively circumscribed written form, and detached from their own directly relevant information needs, while the scientist as reader and citer makes the assessment along the same criteria with regard to the direct present or possible future use for their own specific research agenda and projects.

Having such a common core of assessment criteria, aligned more or less closely with idealistic scientific norms, it is understandable that qualitative peer review and quantitative bibliometric assessment methods have some intriguing parallels. Just one well-known example of such a parallel will have to suffice at this point. In peer review, great care is taken to not select any referee who might have ulterior motives. Thus, family members and partners of assessed persons are usually not eligible, or, if already members of an assessment panel, are asked

to recuse themselves in actual or merely potential cases of conflicts of interest. The parallel in bibliometric practice is the notion of excluding author-self citations (or other types of self-citation, e.g. of journals, organizations, countries) from citation counts to rule out any bias arising from differential rates of self-promotion.

Bibliometrics processes and aggregates these citations as public traces of informal, hidden, private judgments. However, and that is its main limitation, contrary to peer review, bibliometrics does not have access to and incorporate the substantive meanings that scientific works are communicating. Wouters (1999, pp. 198–204) describes this dependency in terms of two information cycles in science: the peer review cycle and the citation cycle, the latter of which latches onto the former but also influences it. Compared to qualitative peer assessment, generic bibliometric citation analysis remains an analysis of formal surface-level features⁷.

Bibliometrics is also intertwined with formal qualitative journal peer review. Ordinary publication counting re-uses the peer review and editorial decisions on manuscripts submitted to journals. This happens when bibliometric studies either select journals or other publication channels (publishing houses of monographs, conference series) based on considerations of the scientific quality standards upheld at these venues or when they rely on the inclusion and exclusion decisions about such venues which are made by the editorial source selection boards of established commercial bibliometric database vendors. Besides inclusion/exclusion selection, a refined approach may also involve the differential weighting of the publication venue in publication counting according to some criterion that may again be based on the collective judgments of the scientific community at large. By these means, fuzzy, not completely reliable, thresholds of discrimination of scientific output by approximate scientific value may be drawn whose numeric values are the results of a great many independent qualitative appraisal events.

In other words, bibliometrics, particularly evaluative bibliometrics, is the scalable analysis of the traces of myriad individual assessment decisions of researchers. In this sense, in one way the quality of bibliometrically informed valuation and decision making is totally dependent on the quality of the under-

⁷By ‘generic bibliometric citation analysis’ I mean citation *counting*. Some access to more nuanced meaning expressed in papers about cited documents is attained by the techniques of citation *context* analysis.

lying micro-valuations: to the degree that the qualitative assessments are made either more conscientiously or more arbitrarily, bibliometric methods which batten onto and aggregate data derived from these decisions gain or lose validity as indicators for impact or quality of research. In another sense, it transcends their idiosyncrasies and the multiplicity of influences of random factors affecting any particular subjective judgments precisely by its scaling and aggregating. It is only by the holistic joint analysis of many publications and citations that bibliometric indicators achieve their utility.

Bibliometric indicator scores are always based on the cognitive processes of a multitude of researchers rather than on those of only the selected few. Bibliometrics thus has intrinsically a much higher cognitive, cultural, and socio-demographic diversity than qualitative peer review exercise can achieve. It also has a different power dynamics. Namely, distributed and pluralistic, rather than the inevitably centralistic and *dirigiste* dynamics of qualitative peer review. And yet these arguably desirable characteristics only come about because bibliometrics is downstream of qualitative cognitive appraisal processes. Attempts at a Manichean juxtaposition of ‘qualitative’ peer review against ‘quantitative’ bibliometrics, of playing one out against the other, can hence be dismissed as either ignorant or misguided or fall under the suspicion of being motivated by vested interests (cf. Abramo, 2024).

1.5.2 Beginnings of bibliometrics

The practice of publication counting with the goal of identifying and studying important researchers and research nations goes back at least to the late 19th century and the *Catalogue of Scientific Papers*. The usefulness and validity of the method was immediately disputed (Csiszar, 2017).

Citation counting became accessible as a technique only with the innovation of the Science Citation Index and this innovation occasioned a momentous shift. The innovation is, as argued by Wouters (1999), the representation of science through scientific literature in a new form. This new form is created by the process of inversion of the cited references into citations. Citations connect a scientific work to the works that use it as a reference. Citations are the elementary objects of citation indexes and they are created by the construction of the citation index. Citation indexes organize citations and make them accessible, objects which apart from a citation index are not realized but would remain po-

tential without it. According to Wouters (1999), the scientist, by referencing a paper, does not give a citation. The citation is created by the citation indexing process that the scientist’s paper undergoes. References are the “raw materials” for citations (Wouters, 1999, p. 112).

Evaluative use of the Science Citation Index began soon after its first publication. Hager et al. (2024) studied the consequences of the publication of the printed SCI in the first years of its existence, the latter half of the 1960s, on U.S. researchers. By exploiting the fact that the early SCI was not published continuously and journal coverage expanded quickly in that period, they are able to make the crucial distinction between citations that were contemporaneously visible in the SCI and thus could have influenced professional researcher careers, and invisible citations, those which were retrospectively added to the SCI and cover the same period but were not available at the time. They show that it was the citations visible in the SCI volumes at that time, but not the invisible ones, that influenced careers. Scientists with more visible SCI citations were better able to secure jobs at higher ranked university departments in their discipline and less likely to drop out of science altogether compared to peers with less visible citations, even given equal numbers of publications. The same is not true for invisible citations. Hager et al. (2024) find that the introduction of the SCI had a large and almost immediate effect on the assortative matching by citation rank between researchers and departments. They also affected promotions to full professor and receiving an NSF grant. This was possible because SCI citation counts, which in the print versions were relatively easy to cumulate by first author names, revealed information on researcher quality that was not possible to obtain before⁸.

Wade (1975) notes several applications of the printed SCI in assessment and research in 1975, including promotion and tenure decisions in universities. According to Garfield (1979, p. 62), “the *SCI* data base is being used to do such things as evaluate the research role of individual journals, scientists, organizations, and communities”. In the U.S., the use of bibliometric indicators for departmental decision-making was well established by the 1990s. Hargens & Schuman (1990) report survey data according to which 35 % of biochemist and

⁸This account of the immediate evaluative use of the SCI is in contrast to Wouters (1999), p.4: “Nevertheless, the scientific community was not enthusiastic. Many researchers did not even use the SCI for its stated purpose as a bibliographic tool – to find relevant publications in the exponentially growing mountain of scientific literature. Neither did many researchers use it to keep abreast of their citation status, a measure without clear meaning to many scientists. The prevailing reaction was hostile or indifferent.”

60 % of sociologist respondents' departments had "made use of citation counts in making decisions about hiring, promotion or salaries." Experimental bibliometric monitoring of research groups began in the 1980s at Leiden University (Moed & van Raan, 1988). The importance of bibliometrics has only increased since then.

1.5.3 The present role of bibliometrics

Currently, metrics whose application on the level of individual researchers are highly controversial in the bibliometric research community, such as the Journal Impact Factor or the h-index, are widely used for academic personnel decision making (Hammarfelt & Rushforth, 2017; McKiernan et al., 2019; Rice et al., 2020). For example, a survey of medical researchers found that the number of publications, first authorships, and the Journal Impact Factor, but not the number of co-authors of publications, were relevant criteria in their performance reviews (Walker et al., 2010). Citation metrics are frequently invoked as relevant evaluation criteria in documents relating to staff review, promotion and tenure in North American universities, albeit sometimes critically (Alperin et al., 2019). A recent Norwegian study has found that for professorial recruitment bibliometrics are used mostly as a first-stage screening tool to reduce the number of candidates for in-depth qualitative review (Reymert, 2021).

While there is support for some application of indicators in personnel decision making among scientists, preference for metrics varies along several dimensions: those researchers who are rated more highly in research quality show stronger metrics preference than those rated relatively poorly; women prefer metrics less than men; STEM researchers are endorsing indicator use more than social science and humanities scholars (Cruz-Castro & Sanz-Menendez, 2021). Attitudes to assessment are then hardly free of motivated reasoning or self-serving biases. As a case in point, the participation in a boycott of a public ranking of German business scholars was shown to be related to how much individual researchers had to gain or lose in such an ongoing change of evaluation culture towards publicly highly visible metrics (Berleemann & Haucap, 2015).

Besides the aforementioned individual researcher assessments, bibliometric indicators are used for all kinds of applications, from use in grant application by review panel members (Gunasekar et al., 2017; Lewison et al., 1999), to national-scale comprehensive research evaluation systems (Hicks, 2012). One

example for the latter is Italy, where the introduction of citation indicators is thought to have led to considerable gaming of bibliometric indicators (Baccini et al., 2019). Performance-based funding systems all over the world rely on bibliometrics. For example, Hammarfelt et al. (2016) found that “almost all major HEIs [higher education institutions] in Sweden use bibliometric measures for resource allocation”. Most German medical faculties use bibliometric measures for funding allocation (Aman & van den Besselaar, 2024). Performance-based funding in Australia on the national level and at universities uses bibliometric data and methods (Woelert & McKenzie, 2018).

There are several national and international research monitoring report series that make extensive use of bibliometrics. Prominent examples are the the U.S. National Science Board’s *Science Indicators*, which started in 1973 (National Science Board, 1973), various reports by the EU, the latest incarnation being *Science, research and innovation performance of the EU, 2024* (European Commission and Directorate-General for Research and Innovation, 2024), and the *Research Fronts* series by the Chinese Academies of Science and Clarivate. Germany uses bibliometric reports for progress monitoring of its major policy instrument for financing extra-university research, the Pact for Research and Innovation⁹.

Perhaps the most publicized use of bibliometric indicators is in global university rankings. All major rankings have been using progressively more sophisticated bibliometric measurement as one component among several others.

This by now ubiquitous metrics-saturated environment has lead to a certain degree of socialization such that use of bibliometric indicators has become a part of scholars’ everyday work, as shown for the social sciences in Australia and Sweden by Haddow & Hammarfelt (2019) or for biomedicine in the Netherlands (Rushforth & de Rijcke, 2015). Haddow & Hammarfelt (2019) characterized this usage as “pragmatism and compliance” (p. 23). Nevertheless, detailed knowledge about the technicalities of indicators is limited among researchers (Kamrani et al., 2021).

Despite fundamental objections, despite frequent inappropriate use, despite limitations pointed out by bibliometricians, despite a surge of both top-down and bottom-up initiatives to rely more on societal impact and non-metric assess-

⁹<https://www.gwk-bonn.de/themen/foerderung-von-ausseruniversitaeren-wissenschaftseinrichtungen/pakt-fuer-forschung-und-innovation>

ments of research – bibliometric indicators have become ever more widespread in application and it looks like they are not going to go away any time soon. While this encroachment and proliferation is frequently objected to, lamented, resisted, and attacked, the data quality of the data basis of metrics-based assessments is surprisingly rarely raised as an argument in these discussions. However, the professional bibliometric research community has in fact devoted some of its attention to this matter. The next section will sketch a selective overview of major topics of this research literature.

1.6 Bibliometric measurement uncertainty as a research subject

Validity and reliability are central issues in the assessment of the usefulness of bibliometric measures for evaluation of scientific performance. (Nederhof, 1988, p. 194)

Validation studies may help to bridge the already existing gap between bibliometric research and technology. Beyond technical validation studies which are designed for detecting systematic and estimating random errors, methodological validation studies may ensure that we really measure what we are intending and assuming to measure. (Glänzel, 1996, p. 171)

population parameters can still be inaccurate due to measurement error, an issue seldom discussed in relation to citation indicators (Schneider, 2013, p. 58)

This section will review research topics in bibliometrics related to uncertainty, data quality, and measurement. It will serve to contextualize the original research of the dissertation. The substantive chapters of this dissertation are all concerned with the role of uncertainty in bibliometric indicators. While uncertainty contingent on data quality is the overarching topic, parts of the dissertation are also concerned with the related topic of measurement validity. One important distinction has to be drawn at this point. Science itself is full of uncertainty. Scientific outcomes of experiments, of research projects, of whole research programs, are uncertain. That is why the research has to be done. For

Leydesdorff (2001), science is essentially the systematic reduction of mankind’s uncertainty about the world. This kind of epistemic uncertainty I must set aside as being outside of the scope of this dissertation – it will only be concerned with measurement uncertainty in bibliometrics.

Known uncertainties are preferable to ignorance about uncertainty. Ignorance about uncertainty can lead to unwarranted belief in the trustworthiness and accuracy of findings, which can lead to misguided decisions. Bibliometricians sometimes call for explicit assessment of uncertainty in results. For instance, in a proposed method for evaluating global university rankings, Gadd et al. (2021), under the rubric of “rigour”, postulate the criterion of *honesty about uncertainty* that rankings should meet: “The types of uncertainty inherent in the methodologies used, and of the data being presented should be described, and where possible, clearly indicated using error bars, confidence intervals or other techniques, without giving a false sense of precision”. Another instance was in the Leiden Manifesto (Hicks et al., 2015), which recommended to “Avoid misplaced concreteness and false precision [...] If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values.” The present dissertation contributes to the development of the understanding, appreciation, quantification, and expression of uncertainty in bibliometrics.

1.6.1 Elements of uncertainty

Uncertainty in scientific results has several distinct components. The first one to consider is sampling uncertainty. Data that is randomly sampled from a population is subject to ordinary sampling error, leading to uncertainty in population parameter estimates. Sampling uncertainty is the subject of the theory of probability and the arsenal of inferential statistics. Bibliometric data often do not constitute a random sample from a larger population. Rather, the data of the population is available or the available data is in no sense a randomly sampled part of the population. In these cases, expressions of statistical sampling uncertainty are unnecessary and inapplicable (cf. Schneider, 2013, for the context of statistical significance testing)¹⁰. When bibliometric data, however,

¹⁰It has been argued that the citation process is probabilistic, that it has stochastic elements (e.g. Nederhof, 1988, sec. 2.2.3), and that therefore the citation count of a publication has some statistical uncertainty, which can possibly be modeled mathematically (Dieks & Chang, 1976). This line of reasoning has not been worked out further nor found application in later

are randomly sampled, the appropriate application of inferential statistics to quantify sampling uncertainty is called for. As an example, in Chapter 2 of this dissertation, I randomly sampled bibliographic items stratified by document type to study document type category assignment error rates. Since these are true random samples of a clearly defined larger population, it was possible and meaningful to calculate probabilistic bounds for the population error rates from the observed sample error rates (confidence intervals). The major source of uncertainty in bibliometric results is usually not statistical sampling. Rather, most of uncertainty in bibliometrics arises from issues of measurement reliability (including data quality) and measurement validity.

We now move to a discussion of the relationships between core concepts of this section. The concept of data quality was already treated in section 1.3. The other concepts are measurement reliability, measurement validity, and measurement uncertainty. As these concepts can be understood to form a chain of dependencies in this specific order, the discussion of their relationships proceeds by that sequence.

In bibliometric research, measurements are made from pre-existing or newly collected data about the scholarly communication system. The quality of this data, primarily its accuracy, may be more or less defective, as will be shown below. The quality of the used data is one factor that affects the *reliability* of measurement. Reliability (of measurement) refers to how accurately and consistently some quantity or attribute can be measured, in other words “the absence of measurement errors” (Alwin, 2005). One useful consideration which is very applicable in bibliometrics is the idea that in measurement there exist both a true score (or value) of a variable, which may only exist theoretically, and that value of the variable which we can observe through measurement, the observed score. The difference between observed and true score is measurement error. A good example from bibliometrics are citation counts of a specific citation database at a specific time. Suppose the measured, or observed, value of a publication’s citation count in a database is 10. We could establish also the true value of that variable with some effort, which is what I did for a sample in Chapter 5. This can be done by manually or semi-automatically checking all similar cited reference strings and counting, independently of the database’s own reference matching procedure, those that refer to the publication in question. Suppose one did this and counted 11 such reference strings. Then the true score is 11

research.

and the measurement error, caused by imperfect data quality, is -1 . Other ways of establishing the reliability of citation counts may be relevant and have been considered. Suppose one is more interested in the reliability of citation counts independent of one specific database. One could then obtain citation counts for the same publications from the same corpus of indexed (citing) literature in different databases and compare them (Franceschini et al., 2013). An issue of reliability of citation counts is that short term citation counts, as collected a few years after publication, correlate more or less with long-term citation counts. Long-term citation counts are arguably the theoretically more appropriate way to measure scientific impact by citations, as the whole citation lifetime of publications is considered rather than just the first few years of citation history. If this reasoning is accepted and it is nevertheless necessary to count citations relatively shortly after publication to have somewhat current data, the reliability of early citation counts for lifetime citation counts becomes an important issue to study (Abramo et al., 2012; J. Wang, 2013).

While it may perhaps appear so in the bibliometric context, data quality and reliability are not the same thing: data quality is one causal factor affecting reliability in bibliometric measurement. Another might be, in the context of the citation count example, the comprehensiveness and quality of the reference matching algorithm.

Another important consideration about reliability to keep in mind is that reliability is only concerned with measured and theoretically measurable values but not with theoretical concepts or constructs. Thus, in bibliometrics, ‘citation count’ is a measurable quantity of a publication record in a citation database. But citation counts are not the same thing as the concept of ‘scientific impact’.

This brings us to *validity*. Validity, in the sense of ‘measurement validity’, refers to how well a specific theoretical concept or construct, which one would like to measure by some measurable quantities, is reflected by these measurements. It is often defined as the “extent to which an indicator of some abstract concept measures what it purports to measure” (Carmines & Woods, 2005). Another way to express the same meaning is to state that validity is the extent to which a measurement instrument’s scores reflect true variation in the construct it is supposed to measure (Whitley & Kite, 2012). Ad hoc measures, those put forward to measure a construct for a study without having their validity demonstrated, “are virtually useless for the assessment of hypothetical constructs” (Whitley & Kite, 2012, p. 174). Science studies investigates many theoretical constructs,

for example scientific collaboration, research quality, interdisciplinarity, prestige, reputation, and knowledge transfer. These are then attempted to be measured by some empirical methods such as surveys or bibliometric methods. For some constructs, validity of measurement instruments can be considered well-established (e.g., for citation counts as partial indicators of research quality, Aksnes et al., 2019, 2023; Thelwall et al., 2023), for others, a validity deficit can be diagnosed (Donner & Blümel, 2024; Q. Wang & Schneider, 2020). This sense of the term validity is to be distinguished from that of the ‘conceptual validity’ of a concept itself, that is, how meaningful a concept is as a description of reality (Locke, 2012).

Reliability and validity are logically independent concepts, but in practical terms, the validity of some measurement of a concept depends on the reliabilities of the variables that are invoked to measure it. Poor reliability of indicators means that a construct, trait, phenomenon, etc., cannot be accurately measured by a measurement process. This is why, especially in bibliometrics, considerations of data quality fall under the rubric of measurement reliability. Consider that a specific kind of data is theoretically a highly relevant indicator for a phenomenon. If it turns out that the data that was collected is affected by many errors, the measurement cannot be reliable. If the indicator cannot be reliably measured, the measurement of the concept by this indicator cannot have high validity.

Systematic measurement error can impede the measurement validity of a measured indicator for a concept. Systematic error in measurement, viz. bias, constitutes non-random variability in the measured values for a concept that is unrelated to the meaning of the concepts. One case of measurement validity in bibliometrics being lower than it could be is treated in Chapter 3: *Effect of publication month on citation impact*, so I will use this case as an example here. If the scientific impact of publications is measured by citation counts as a proxy variable, or indicator, then systematic variation in the indicator unrelated to the concept of impact should be eliminated as much as possible. Doing so is standard practice in professional bibliometrics and is usually referred to as normalization. It is not always made explicit that the reason for normalization is to remove variability systematically unrelated to a specific measured concept. It is more common to find the justification that one would like to make raw citation counts more comparable to each other, which is a more non-technical expression of the same concern. Citation counts are routinely normalized for several char-

acteristics which are argued to distort the relationship between measured values and the concept of scientific impact. For instance, if publications published at different times have to be compared, it is arguably necessary to normalize citation counts for the effects of longer or shorter duration of being published and available for citation because the citation counts of publications published earlier will be systematically higher. A difference which is unrelated to different levels in scientific impact. In Chapter 3, this argument is developed further because the standard methodology for citation count normalization only normalizes for publication year. In that chapter it is argued and demonstrated that a citation count bias also exists in favor of publications which were published earlier over those published later in the same year. It is then demonstrated that this systematic error can be eliminated by using month-precision citation windows, as opposed to the conventional year-precision citation windows. Month precision citation counts thus have higher validity for the concept of scientific impact because they remove bias in the measured variable.

Both reliability and validity are causes of uncertainty about the results of measuring a concept. Measurement uncertainty refers to how uncertain (imprecisely known) the value of an obtained measurement is, given all available information about the measurement process, including random and systematic error. The *Guide to the expression of uncertainty in measurement*, an authoritative document for measurement uncertainty in the physical sciences, defines the concept as follows: “parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand” (JCGM, 2008, p. 2), where “measurand” refers to the quantity that is to be measured. The *Guide* notes that this definition is compatible with uncertainty as a “measure of the possible error in the estimated value of the measurand as provided by the result of a measurement; an estimate characterizing the range of values within which the true value of a measurand lies” (p. 3). Measurement uncertainty is hence a quantified expression on how precisely a measurement reflects the underlying but unobservable magnitude of a concept taking into account all pertinent knowledge about all sources of uncertainty in the process of measurement.

1.6.2 A panorama of sources of uncertainty in bibliometrics

We can now continue with a discussion of the variety of sources of bibliometric measurement uncertainty that have been reported in the literature. The following discussion does not attempt to be exhaustive or representative but illustrative. The point is to demonstrate the wide range of primarily data quality issues that cause measurement uncertainty.

1.6.2.1 Accuracy, Completeness, Duplicates It is probably safe to say that all bibliographic data are affected more or less frequently by inaccuracies. Author names can appear incorrectly on publications' author lists due to mistakes in proofing by publishers (Teixeira da Silva, 2024). Demetrescu et al. (2018) compared author name form in four major bibliographic data sources with canonical variants for a large number of mainly Italian names. They find error rates (incompatible name strings) between 0.25 and 3 %. Not only authors' names, but also the information on corresponding authorship has been found to be affected by substantial inaccuracy in bibliometric databases (Chinchilla-Rodríguez et al., 2024).

DOIs in bibliometric databases are affected by errors of different types. For example, one DOI can be assigned incorrectly to more than one item or an item can be indexed with the wrong DOI (Franceschini et al., 2015a; Krauskopf & Salgado, 2023; Liu et al., 2021; Xu et al., 2019; Zhu et al., 2019). Franceschini et al. (2016) discuss several other identified data quality problems in Scopus, among them spelling mistakes in article titles, the overlooking of citation links to papers cited while available online without issue and page information, duplicate records for the same publications, cited references in the database record that are not present in the original paper, and cited references missing most of their information. Inaccurate funding data has also been reported for the major databases (Álvarez-Bornstein et al., 2017; Grassano et al., 2017; Liu, 2020). It has been found that WoS has not indexed author affiliation data that is present in the original publications (Liu et al., 2018). Commercial bibliographic-bibliometric databases have been found unknowingly to index content by fraudulent, hijacked journals (Abalkina, 2024).

One chapter in this dissertation, with the title *Document type assignment accuracy in the journal citation index data of Web of Science*, is an investigation into the prevalence of another kind of incorrect or inaccurate values. There I

study the extent to which the document type categories assigned to a sample of publications during indexing in citation databases differ from an independent intellectual assignment of document types using the definitions and criteria of the citation database. Another earlier study in which the correctness of document types played a role is that of Heneberg (2013) on seemingly uncited papers of Nobel Prize and Fields Medal laureates. Heneberg found that among the records in WoS of these laureates which were both uncited and categorized as articles or reviews, 26 % were incorrectly categorized and were actually of different document types.

Inaccurate citation counts and errors in cited references lists are a major topic of the bibliometric data quality literature (e.g., Cioffi et al., 2022; Franceschini et al., 2013; Liang et al., 2014; Moed & Vriens, 1989; Olensky, 2015; Olensky et al., 2016; Tüür-Fröhlich, 2016; Xu et al., 2019). The accuracy of WoS citation counts is also a component of the Chapter *Data inaccuracy quantification and uncertainty propagation for bibliometric indicators*. This chapter also gives a thorough overview of the literature on this topic, so it will not be preempted here.

Occasionally, inaccurate citation counts come under the spotlight of attention, such as when *Nature* staff noticed an incorrectly low Web of Science citation count of a landmark human genome paper published in their journal and wrote the following in an opinion piece:

When puzzled Nature staff examined the ISI's data, they found that citations to the paper had been grossly undercounted. The same, it emerged, applied to several other prominent papers authored by a consortium, rather than by a list of individuals. To its credit, the ISI reacted promptly, amending its website and commissioning a journalist to write an account of the episode for Science Watch. The paper by the International Human Genome Sequencing Consortium now sits at the top of the hot papers list. (Anonymous, 2002)

It is worth noting in this context that papers with group authors are particularly susceptible to indexing and citation errors (Dickersin et al., 2002).

The above mentioned study, Heneberg (2013), identified further causes of apparently uncited publications of Fields and Nobel laureates besides mis-classified document types. Of the 100 ‘uncited’ articles and reviews (according to WoS),

Heneberg was able to find citations by manual cited reference search for 46 of them, which were simply not linked by WoS due to errors present in the original citing document or introduced by WoS indexing¹¹.

When authors cite references with incorrect information, matching the cited reference information to the correct intended target becomes more difficult. The unique study of Waytowich et al. (2006) showed that the accuracy of cited reference data is related to individual differences in psychological traits. They studied PhD candidates' referencing errors and deviations from a proscribed citation style norm in dissertation proposals and these error scores' relationships to a measure of the construct of 'perfectionism', to expectations of study success, and to demographic variables. The two citation list quality scales were found to be correlated. Students with lower course success expectations made more citation mistakes and students with more completed courses showed lower adherence to citation style rules. Analyzing citation quality and perfectionism scales showed that "students who reported the lowest levels of self-oriented perfectionism and socially prescribed perfectionism tended to commit the most citation errors and constructed reference lists that departed the furthest from APA (2001) stipulations" (Waytowich et al., 2006, p. 203). While this study must be interpreted with some caution due to small sample size, it would, if confirmed by more evidence, point to the hypothesis that at the level of researchers committing referencing errors, stable processes are at work that are connected to the individual dispositions of persons.

In a fine piece of investigative bibliometric detective work, Besançon et al. (2024) have recently documented a novel manipulative phenomenon which they have dubbed 'sneaked references'. Sneaked references are the intentionally disguised counterpart to missing references. They are references which are not present in the original publication but inserted into their metadata at the registration of this metadata at Crossref and from there on propagated to information systems. The patterns uncovered by Besançon et al. (2024) indicate intentional exploitation of this system for nefarious purposes. From this example it is apparent that true data quality issues, here the inaccuracy of values, need not be caused by some error-prone processing step or negligence. They can be caused by systematic abuse of infrastructure. In this case they lead to increases in citation

¹¹This finding, incidentally, points to the relative fragility of indicators of uncitedness. Another reason to be skeptical about uncitedness indicators, also deriving from their dichotomous nature, is that is probably the easiest citation indicator type to influence by self-citation.

counts of cited papers, the cited references of which are hidden.

1.6.2.2 Entity disambiguation, reconciliation, matching, deduplication

Various entities that are treated as actors in the socio-cognitive system of scientific research are not readily identifiable by looking only at their names. Entities such as authors, research performing and funding organizations (and their subdivisions), countries (and their subdivisions), journals and other publication venues, and publishers all appear in the written documentation of scientific research with various different surface forms of their names. This includes, of course, alternative names, changed names, abbreviations, translations and transliterations, spelling mistakes, and probably more. Different persons and organizations may happen to share a name or name form. All these ambiguities, only some of which are data errors, have to be resolved by linking the various surface forms of an entity to its identity or a preferred form, a process that information science describes by the somewhat ominous term of authority control and that computer science calls (entity) disambiguation or reconciliation. Entity disambiguation systems, usually algorithms, are themselves a source of inaccuracy since none of them work with perfect precision. The performance of such systems is conventionally judged by their level of accuracy.

The disambiguation of the names of all of these mentioned entities has been the subject of bibliometric research (Leydesdorff et al., 2016; Prancutè, 2021). In particular, the disambiguation of author names and of organization names have attracted a lot of attention, also from computer science, because of their obvious practical importance.

To give just one example, Sinatra et al. (2016) studied the timing of researchers' most impactful publication in their careers and the influence of individual level factors on citation impact. For author name disambiguation, they iteratively merge authors in an American Physical Society data set based on simple criteria, one of which is identical family name, meaning that the records for people who have changed names will never be unified. They explicitly report false positive and false negative rates of intellectually checked samples. The easy availability of this data set has prompted a fair amount of re-use. Such heuristic author name disambiguation approaches are relatively scalable while resulting in reasonably good disambiguation quality. Similar approaches are, for instance, D'Angelo et al. (2011), Caron & van Eck (2014), D'Angelo & van Eck (2020). The results of studies with disambiguated author names are systematically different to the

otherwise identical studies using unprocessed names (Erman & Todorovski, 2015; Schulz, 2016). In other words, bibliometric research which studies authors as its object and uses undisambiguated name data will be systematically biased against some authors or groups of authors and their results cannot be accepted at face value.

The names of research performing organizations, too, have to be unified if one wants to carry out meaningful research in which the objects of study are organizations. The question of disambiguation of the identity of research organizations from address-affiliation data is taken up in Chapter 4, titled *Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems*. The study compares the disambiguation results of different systems on the same bibliometric data from a perspective of applied bibliometrics. This chapter also provides an overview of related work on institutional disambiguation for science studies. A related study is that of Orduña-Malea et al. (2017), who provide an empirical study and assessment of the quality of Google Scholar’s affiliation feature – an example for the relatively rare studies of Google Scholar’s data quality. They identify several problems in the unification processes that lead to, for example, universities with similar names being conflated under one affiliation and universities being subsumed under the wrong normalized name.

Bibliometric databases or analytical data sets may comprise records from more than one external data source. If the coverage of the sources overlaps, there will be duplicate records in the derived data that need to be identified and deduplicated, that is, all but one record removed or all records consolidated into one. Valderrama-Zurián et al. (2015) studied duplicate records in Scopus. They were able to attribute such duplicates to mis-mapping of records to more than one journal, journal title changes and spelling variants, and journal name variants – these latter resulting from indexing records from multiple data sources with insufficient deduplication. Liu et al. (2021) studied indexation errors in WoS and Scopus of one particular megajournal and identified record duplicates as one type of error.

More generally, whenever data sources are combined to satisfy information needs, one faces problems of data integration and interoperability. One comprehensive approach to deal with such issues of data governance and management that has also been used for implementations geared towards evaluative and monitoring purposes for supporting science and innovation policy is ontology-based

data integration. This is a top-down approach in which the domain of knowledge is formally modeled in a machine-readable way and mappings from the ontology to concrete data sources are created. Crucial advantages of this approach, according to Lenzerini & Daraio (2019), are the ability to logically reason over the domain concept structure, the mapping, and the data. Besides that, this approach has provisions for data consistency and data quality analysis.

1.6.2.3 Completeness Data that is present in the original publications is sometimes found to be missing in the indexed records of these publications in bibliometric databases. Jacsó (2009) provided an early comparison of WoS and Scopus specifically pertaining to missing values in essential data fields. He finds a considerable missing rate for document type in Scopus, noting that “This is important to know because using the document type to filter the search results could exclude millions of wanted documents that do not have any document type information. The nine per cent omission rate [of Scopus] is too high for scientometric purposes”. The other important difference between the two data sources was the much higher rate of missing country information in Scopus, although the reasons for this could not be determined. The ramifications are nevertheless obvious for bibliometric studies using this data source: “Irrespective of the reasons for the very high rate of omission of country names or codes, it should be realised and prominently mentioned in any scientometric country reports. I have never seen this mentioned in published papers, nor in the manuscripts that I have peer reviewed.” Liu et al. (2021) reported dozens of missing records for major bibliometric databases for a single large journal.

For bibliometric studies it is often important to have not only the author names or identities and the organization names and identities which are responsible for publications but also the information about which authors are associated with which organization and country – the author-affiliation links. This data can be useful, for instance, to estimate the degree of contribution of co-publishing organizations in jointly produced work. Web of Science has indexed these author-affiliation links only from 2008 on (Birkle et al., 2020). This means that studies using WoS data that need this information cannot calculate correct values for the period before 2008.

Missing references are cited references present in the original publication but erroneously not included in the information system or citation database. They have been repeatedly observed (Buchanan, 2006; García-Pérez, 2015; Rodrigues

et al., 2023; Van Eck & Waltman, 2017).

Curiously, major commercial databases do not index some information that is useful both for information search and bibliometrics which are routinely available in published papers, such as the editorial handling dates (received, reviewed, revised, accepted dates) and the classification codes of various disciplinary classification systems. But it cannot be said that this constitutes missing information since there was never an intention to include these data.

1.6.2.4 Methodological robustness and sensitivity analysis Robustness tests and sensitivity analysis are tools to ensure that the results of a study are not artifactual and vanish when minor methodological variations are applied. They concern not only individual values of variables but the whole pattern of results of studies. Robustness checks consist of more or less systematic tests of whether findings can be reproduced under alternative operationalizations that should not overturn the findings qualitatively, such as using a different but equally applicable data source, a different population of study, alternative measurement instruments, alternative statistical techniques, etc. Sensitivity analysis concerns more specifically the robustness of results to systematic variations in the values of free parameters of algorithms that researchers have to set. Findings that remain invariant or qualitatively unchanged under such changed conditions are evidently robust to alternative specifications of a research methodology and can be said to be relatively more epistemically reliable compared to findings that do not survive such tests. When the robustness checks concern specific values and their robustness under changed conditions within a given methodology, such as variations in parameter values of algorithms or the criteria for selection of objects of observations, there will be ranges of result values for result parameters of interest. These ranges can be interpreted as the uncertainty in result values that can be attributed to minor variations in methodology (cf. Stahlschmidt & Schmidt, 2019).

The Leiden Manifesto (Hicks et al., 2015) explicitly suggested such checks: “Science and technology indicators are prone to conceptual ambiguity and uncertainty and require strong assumptions that are not universally accepted. The meaning of citation counts, for example, has long been debated. Thus, best practice uses multiple indicators to provide a more robust and pluralistic picture.” Especially when there is insufficient knowledge about the comparative validity of alternative analytical choices, one way of addressing this is sensitivity analy-

sis. However, this should not be confused with the uncertainty in results that can be derived from known validity coefficients of specific applied instruments, indicators and other methods.

A paradigmatic case is the debate on the alleged relative decline of British science in the 1970s-80s. Irvine et al. (1985) and B. R. Martin et al. (1987) opened the debate, finding a decline in the percentage of British science of the world total according to figures computed from CHI's database compiled from SCI data. The CHI databases used fixed journal sets and the methodology involved fractional counting of papers at the level of countries. The central finding, a decline in British world percentage of paper production, was challenged by Leydesdorff (1988), who reproduced the original findings but contrasted them with results showing an increase in Britain's world contribution when using the dynamic journal set of an online version of the SCI and whole counting. Kealey (1991) also criticized the journal set albeit without presenting evidence, but tried to change the debate completely by postulating that per capita numbers of papers and contribution to the couple of hundreds of most cited papers per year are what matters. These indicators presented a much more favorable view of the status of British science. Braun et al. (1989a) went some steps further and did an actual sensitivity analysis by processing the same data with different justifiable methodological choices, including journal set specification and counting method, but also others. They found "the annual mean relative change of the 28 indicators varied between -0.88% and $+4.30\%$ with an average of $+0.93\%$. Since the standard deviations within each time series were in the range 1% to 3%, no statistically significant trend could be attributed to any of the indicators." The substantive and methodological debate by no means ended there, it was extensive and detailed and continued for years (Leydesdorff, 2017; B. Martin, 1991, 1994; B. R. Martin, 2017). As this first example illustrates, methodological choices can and do have important ramifications for bibliometric findings and great care must be taken that results are robust to minor changes in methods. Even in the absence of empirical validity knowledge about methods, trying out different methods and parameters can tell us if these choices make a difference or not. This explanation makes it clear that robustness/sensitivity analysis does not neatly fit into the conceptual sequence (data quality $\rightarrow$ reliability $\rightarrow$ validity $\rightarrow$ uncertainty) of this dissertation. One may conceptualize it as a distinct side chain that itself on one side ends in uncertainty: (methodological degrees of freedom $\rightarrow$ robustness/sensitivity analysis $\rightarrow$ uncertainty $\leftarrow$...). As robustness

and sensitivity analysis are important tools to assess the overall uncertainty of results I have discussed them here. They are, however, not treated further in the substantial chapters of original research that follow.

As a more recent example of a dedicated systematic sensitivity analysis, Stahlschmidt & Stephen (2022) investigated how the different scopes of content indexation in Web of Science, Scopus, and Dimensions affect citation impact analysis in bibliometrics. By calculating normalized citation scores for publications with German affiliations indexed in the three sources, the study finds that it does make quite a difference which citation data source is chosen for the measured citation impact performance of German science system sectors. Another example for a dedicated study of the possible factors influencing results in bibliometrics is that of Struck (2025), which studied a multiplicity of causes that affect the results of citation network-based clustering of publications used for detecting and tracking topics. Some of the identified dependencies are the data source, software configuration, data model, and clustering algorithm. Another clear validity deficit is diagnosed in this area.

A crude but common method of checking for methodological robustness with regard to the influence of different methods of accounting for co-authorship in publication counting, i.e. credit allocation methods, is to run the same analysis but using different such methods and assessing if the findings remain stable. Porter (1977) (p. 260) was an early advocate of sensitivity analysis in science studies in the context of methodology of allocating credit to co-authors of multi-author papers. One recent example is Sinatra et al. (2016), who used equal fractional counting, an author order-sensitive allocation method, and a collective perception method of attributing publication credit to co-authors. In addition, they also repeat the main analysis only for authors whose most-cited papers are solo-authored, in order to rule out confounding by co-authorship in general. The robustness checks of Sinatra et al. (2016) go much beyond counting methods. They test and ensure that their main findings reproduce under many different conditions, including removing difficult to disambiguate names entirely, using different base data sets, different choices of selection of observed authors in these data sets, rescaling key citation metrics, and excluding review papers. Another study with comprehensive robustness checks is that by Hager et al. (2024), which was mentioned before in the context of the early adoption of the Science Citation Index for faculty review and hiring. The application of robustness checks in scientometric research has been raised to a new

higher standard by the introduction of multiverse analysis, a rigorous systematic approach to methodological robustness in empirical research, by Leibel & Bornmann (2026).

1.7 Organization of the dissertation

What follows are the four chapters of original research that together comprise the research content of this dissertation. The general guiding research question of this research project is:

How does data quality affect the reliability of bibliometric indicators?

The chapters are structured and organized as follows.

In Chapter 2, I study the prevalence of inaccuracies in the indexed document types of publications in citation databases and whether these mis-assignments have a measurably distorting influence on the calculation of normalized citation impact indicators. The accuracy of document type classification in scholarly information systems has never been studied before and the same holds for consequences of inaccurate document type data on citation indicators. For this study, a random sample of records was drawn from WoS, which was stratified according to WoS document type. This means that ordinary statistical inference is applicable and is used in the form of population parameter estimates with confidence intervals. Document type accuracy is assessed by manually re-classifying this sample's document types according to the database's definitions and then comparing the independent manual document types assignments to those of the database. Incorrect document type classifications will lead to the inadvertent exclusion or inclusion of unwanted publications with wrongly assigned categories when only publications of certain types are selected for a study, which is a very common step in bibliometric research and assessment (Jacsó, 2009; Waltman, 2016, p. 372). I further study if the document type classes assigned to the same papers by the two databases Web of Science and Scopus are consistent. Moreover, I also study in which way and to which magnitude inaccurate document type classification exerts an influence on normalized citation scores, which are a core element of advanced citation analyses. As normalization is generally

carried out with respect to the variables publication year, scientific discipline, and document type, some influence of document type errors is expected.

In Chapter 3, another issue related to citation count normalization is studied, namely the dependency of citation counts on publication month. The conventional method of normalization for time in citation analysis is to compare only the citation counts of papers published in the same year. The reason is that papers published in earlier years than others have had more time to accrue citations. But this problem also exists for papers published in the same year: there are large differences in average citation counts across publication months to the advantage of earlier over later months (as pointed out before, e.g., Gai et al. (2015)). In this chapter I investigate this effect with a large-scale systematic study. This includes the existence of this effect (or bias), its size, and a test of a method to remove it. As citation month data is not always available in citation index databases I also test and evaluate a method to infer this information from other publication data. This chapter is therefore concerned with the reliability and validity of citation counts and normalized citation scores as indicators of scientific impact. If the factor publication month has a systematic distorting effect on citation counts which is unrelated to the construct that citation counts are used as a proxy for, then this bias should be removed to improve the measurement validity of citation counts. Finally, I assess if the publication month effect matters in a research evaluation scenario. To this end, I test in the chapter if aggregated citation scores of German research organizations are measurably affected by publication month bias.

In Chapter 4, the issue of institution affiliation disambiguation is investigated. In this co-authored study, we first empirically establish the very high accuracy of a reference data set of records on publications from Germany. We proceed to use this data as a gold standard to compare the quality of the affiliation disambiguation systems of both Web of Science and Scopus, plus the results of using institution name searches. We analyze the performance of the affiliation disambiguation systems in retrieving the publication sets of research performing institutions, as far as covered in the respective databases, as completely and accurately as possible. This chapter also includes an analysis of the real-world effects of the possibly insufficient data quality of the disambiguation systems on a broad set of standard bibliometric indicators from the three domains of publication output, scientific collaboration and scientific impact. This chapter is thus directly concerned with the pathway from data quality to reliability of

bibliometric indicators.

Chapter 5, in a first part, continues the empirical study of bibliometric data quality by looking at WoS citation counts. The citation counts calculated by WoS are compared to citation counts derived from semi-automatic cited reference searches, that is, manually collected gold standard data. This chapter is therefore also concerned with the influence of data quality on the reliability of bibliometric indicators. In the second part, a new approach to the propagation of uncertainty from data quality to indicator values is introduced. This approach builds on Bayesian modelling of error distributions and posterior prediction of simulated error-free values. I show how this approach can use knowledge on statistics of inaccuracies from empirically collected samples to generate error margins for the values of bibliometric indicators of interest. The method is applied using both the citation count error data from the first part of the chapter and the data on document type misclassification from Chapter 2. It is noteworthy that this method can in principle be used for propagation of errors from any bibliometrically relevant data. The usefulness of this method in real bibliometric evaluations is shown for data from an evaluation of the research output of chemistry research groups. In this case study, I also use the month-precision citation windows from Chapter 3. This chapter builds on the results of the earlier ones but goes beyond them in sketching a method which allows to use the many empirical results on data inaccuracies to construct errors margins for bibliometric indicator values, an important demand of the Leiden Manifesto (principle 8).

In these studies, a number of different issues related to reliability and validity of bibliometrics are studied which are primarily related to data quality. In some cases it can be shown that the processing of the original data for the bibliometric databases is the location where inaccuracies are introduced (document type classification), in some cases data is missing that is in principle available (publication month data), in other cases the original data are heterogeneous, or unharmonized, and the processing by the bibliometric information systems only manages to resolve some, but not all, of this heterogeneity (cited reference strings, address affiliation strings). The common approach to quantifying the data quality in the bibliometric information systems is the direct empirical comparison with outside reference records. Besides identifying, quantifying, and in some instances demonstrating methods of remediating such errors and inaccuracies, one central concern of these studies is to investigate the effects of these

data quality issues on bibliometric practice in basic and applied research. The intention is to show that these are not mere technicalities or errors so small that they vanish in large scale data analysis. I am confident that after having read the next few chapters, the open-minded reader will agree that these problems do actually matter a great deal, that they deserve to be taken seriously and that the proposed solutions can contribute to improving the professional practice of science studies.