



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).

English language summary

Bibliometric uncertainty: Data quality, reliability, and validity

Bibliometric methods and bibliometric indicators are important, albeit disputed, tools in empirical studies of science systems and in research evaluation. Their validity and acceptance is partly due to their statistical scope, allowing for analyses at large scales, such as entire countries, for which the cost of employing qualitative methods is too high. Bibliometric methods are deeply data-driven which makes the trustworthiness of the statements and interpretations derived from bibliometric studies very data-dependent. Science studies researchers have long recognized that poor quality data is a possible threat to informative scientific outcomes and solid, defensible quantitative assessment of research. They have consequently put considerable effort into empirically investigating the quality of their data although this has not yet precipitated the emergence of a consolidated systematic research program in its own right. Scientometricians and bibliometricians have studied such data quality topics and author name disambiguation, organization name disambiguation, completeness and accuracy of various bibliographic fields such as cited references, identifiers (DOIs), and assigned document types and categories of subject classification systems. Not infrequently, the practical consequences of the observed error rates have remained unclear. How to express the amount of uncertainty in bibliometric that are caused by inaccuracies in the basic data has also remained elusive. The research of this dissertation extends the knowledge on error rates of specific data attributes, namely document types and citation rates, and studies also the performance of

different organization disambiguation systems. It further examines the standard methodology of citation analysis of using nominal year-based citation windows and find this to be biased. In all studies, emphasis is put on the real-world implications of these deficiencies and possible improvements. The final part introduces a novel framework for the propagation and reporting of uncertainties that arise from such data quality issues. I show how inaccuracies in input data can be used to derive realistic error margins for bibliometric indicator result values that quantify their uncertainty due to statistical error in the data. The main result is that bibliometric indicator values are commonly affected by the error sources that were studied that small to moderate differences in indicator values—even for aggregate units such as research groups, departments, universities—might be attributed to data inaccuracies rather than actual performance differences. The implications and consequences of these findings are discussed and recommendations for improvements of current practice are derived.

Nederlandstalige samenvatting

Bibliometrische onzekerheid: gegevenskwaliteit, betrouwbaarheid en validiteit

Bibliometrische methoden en bibliometrische indicatoren zijn belangrijke, zij het omstreden, hulpmiddelen in empirische studies van wetenschapssystemen en in de evaluatie van onderzoek. Hun validiteit en acceptatie zijn deels te danken aan hun statistische reikwijdte, waardoor analyses op grote schaal mogelijk zijn, zoals op nationaal niveau, waarvoor de kosten van het gebruik van kwalitatieve methoden te hoog zouden zijn. Bibliometrische methoden zijn sterk datagedreven. De betrouwbaarheid van de uitspraken en interpretaties die uit bibliometrische studies voortvloeien zijn dan ook sterk afhankelijk van data. Wetenschappers erkennen al lang dat data van slechte kwaliteit een bedreiging vormen voor informatieve wetenschappelijke resultaten en een solide, verdedigbare kwantitatieve beoordeling van onderzoek. Ze hebben daarom aanzienlijke inspanningen geleverd om de kwaliteit van hun data empirisch te onderzoeken. Dit heeft echter nog niet geleid tot de opkomst van een systematisch onderzoeksprogramma. Scientometristen en bibliometristen hebben de datakwaliteit bestudeerd, evenals het doorverwijzen van auteursnamen, organisatienamen, de volledigheid en nauwkeurigheid van verschillende bibliografische velden zoals

geciteerde referenties, identifiers (DOI's) en toegewezen documenttypen en de gehanteerde categorieën van classificatiesystemen van onderwerpen.

Niet zelden zijn de praktische gevolgen van de waargenomen foutpercentages onduidelijk gebleven. Hoe de mate van onzekerheid in bibliometrische gegevens, veroorzaakt door onnauwkeurigheden in de basisgegevens, moet worden uitgedrukt, is eveneens ongrijpbaar gebleven.

Het onderzoek in dit proefschrift breidt de kennis uit over foutpercentages van specifieke dataattributen, namelijk documenttypen en citatiepercentages, en bestudeert ook de prestaties van verschillende systemen voor het ondubbelzinnig maken van de verwijzingen naar organisaties. Het onderzoekt verder de standaardmethodologie voor citatieanalyse met behulp van nominale jaargebaseerde citatievensters en concludeert dat deze bevooroordeeld is. In alle studies wordt de nadruk gelegd op de praktische implicaties van deze tekortkomingen en mogelijke verbeteringen.

Het laatste deel introduceert een nieuw raamwerk voor de verspreiding en rapportage van onzekerheden die voortvloeien uit dergelijke problemen met de datakwaliteit. Het proefschrift laat zien hoe onnauwkeurigheden in invoergegevens kunnen worden gebruikt om realistische foutenmarges af te leiden voor bibliometrische indicatorresultaatwaarden.

Het belangrijkste resultaat is dat bibliometrische indicatorwaarden vaak worden beïnvloed door de onderzochte foutenbronnen, waardoor kleine tot matige verschillen in indicatorwaarden – zelfs voor geaggregeerde eenheden zoals onderzoeksgroepen, afdelingen en universiteiten – mogelijk te wijten zijn aan onnauwkeurigheden in de gegevens in plaats van aan daadwerkelijke prestatieverschillen. De implicaties en gevolgen van deze bevindingen worden besproken en er worden aanbevelingen gedaan voor verbeteringen vande huidige praktijk.