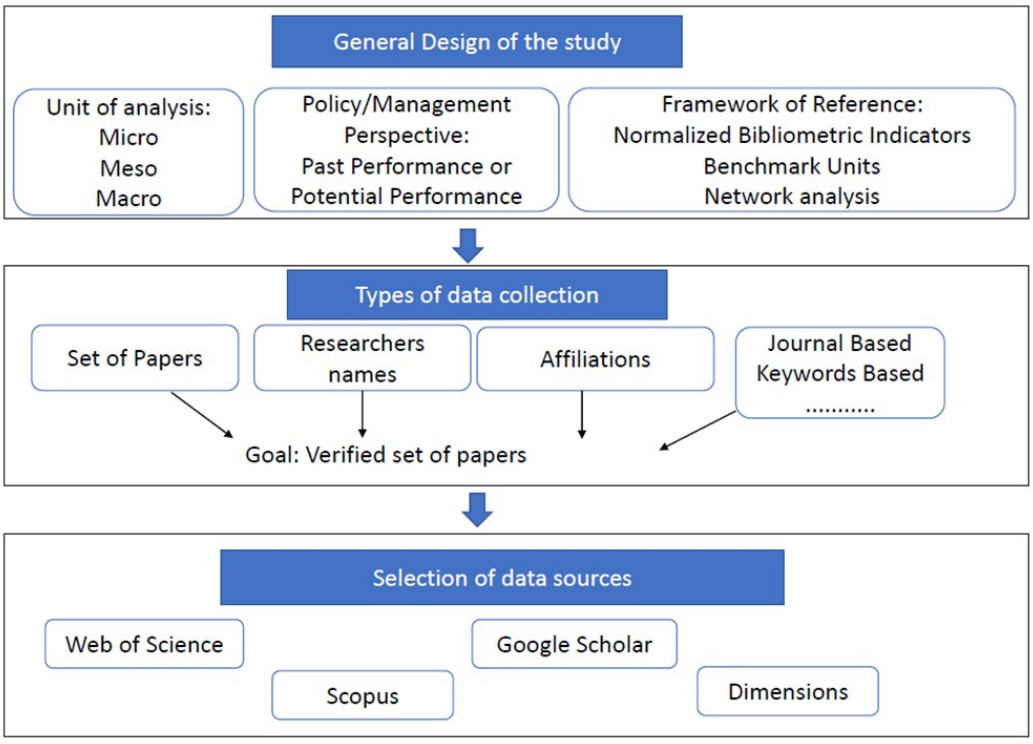
Data sources and data collection

Publications are one of the main outputs of research activities, and various benchmarking surveys and surveys measuring the level and impact of science based on publishing activities are well established in the scientific community. While such surveys undoubtedly have their place in the Finnish scientific field, it is important to remember that not all methods of publication metrics are directly applicable to all levels of analysis. It is therefore necessary to distinguish between different approaches to publication metrics and to understand the results of the chosen methods. If publication-based analyses are to be included in the evaluations of research activities, it is important to be able to select the appropriate methods and tools according to the need for use.



Graph 1. The planning, data collection and data sources of publication-based analysis. © Anna-Kaisa Hyrkkänen, 2022 (based on the CWTS Leiden course material *Measuring Science and Research Performance* by Visser M. and Calero C., 21 September 2015)

General design of the study

The unit of analysis

The scope of the unit under analysis can range from a very broad country-level study down to individual level. In the context of publication-based analyses, we typically talk about **micro, meso and macro levels**. Typical micro-level analyses are those based on the publication activity of an individual researcher or an individual research group. Finding out a researcher's citation numbers from a particular database is an example of a micro-level analysis. At the meso level, for example, subject of the analyse can be a single research organisation. Research evaluations in universities and related publication-based surveys are good examples of meso-level analyses. At the macro level, an even broader picture is examined, for example by analysing the publication activities of an entire country. The State of scientific research in Finland reports carried out by the Academy of Finland are an example of macro-level analyses. The chapter Targets of analysis explains in more detail how different units can be examined using methods of publication metrics.

Pay attention to responsibility, especially in micro-level analyses! The unit of analysis is directly related to the risks associated with publication-based analysis. The lower the level of analysis, the higher the risks associated with its use. An analysis can have a greater impact at individual level than, for example, at organisation level. Remember that publication-based metrics can be used to support a qualitative expert evaluation, but it must not be the only method for evaluating universities, their units or individual researchers in particular.

Policy/management perspective: Past performance or Potential performance?

If the purpose of publication metrics is to measure scientific impact, it is important to distinguish between two different perspectives. Is the purpose to increase understanding of the past performance of the unit under analysis by means of retrospective analysis or to gain insight into the research potential of a particular unit by means of prospective analysis? The perspective of the review largely determines which data collection method is used. As a practical example, we can think about the research evaluation carried out in a research organisation. The research data can be collected by extracting the publications affiliated to the unit under review from the citation database for the chosen time period. Typically, the time period is chosen so that publications have at least two full years to receive citations. In such cases, the review is a retrospective analysis of the research citation impact. Alternatively, the research data can be collected by extracting the publications produced by a specific group of people, such as professors at the unit under review, from the citation database. In such cases, the analysis focuses on the total publications by the selected individuals over the chosen period, regardless of the organisation under which the publications were produced. This latter method of data collection allows the future research potential of the unit under review to be examined, which means that it is a prospective analysis. It is also possible to combine these two data collection methods during the data collection phase.

Framework of reference: What does the analysis aim to find out?

The methods used in publication metrics make it possible to analyse the unit under review from different perspectives, the most typical of which are:

- Surveys that measure the scope, level and impact of research, which can be at micro, meso or macro level. By examining the number of publications, it is possible to determine which organisations carry out research in a certain field of science and how extensive the research activities are. The level and impact of research can be approximated using normalised citation impact indicators. For example, they can be used to determine the proportion of a unit's publications that are among the most cited publications in the world.
- Benchmarking surveys (comparative analysis), which in turn compare the performance of the unit under review, such as an individual researcher, a research institution, a university or a country, with other similar units. In the interest of responsible metrics, it is important that the units selected for review are comparable with each other. There is no reason to compare the research or researchers in different fields of science through publications, as there are significant differences in publication and citation practices between different fields of science.
- Network analyses, which can be used to examine relationships between publications based on the authors of the publication, terms used or citations. Collaboration networks can be used to examine the collaboration between the authors of publications or organisations. Semantic network or co-word analyses can be used to analyse the most frequently occurring words in a certain set of articles. These kinds of analyses are useful for looking at how different fields of research are linked or what similarities there are between them. Citation network analyses examine the relationships between publications through citations. Citation relationships can be direct, in which case the publications under review cite each other, or they can be analysed through co-citations, looking either at articles that cite the same publications, or at articles that are cited by the same publications.

Types of data collection

In order to be able to analyse the publication activities of a unit, the publications produced by the unit must first be reliably identified. The data can be collected, for example:

- By collecting a set of publications: for example, the publications of the unit under review can be extracted from the organisation's research information system for the desired time period.
- By the names of researchers: by defining the researchers whose publications constitute the data under review. For example, if you wish to examine the publication activities of a particular researcher group, you should start by listing the researchers whose publications you want to include in the analysis. During the data collection phase, care should be taken to ensure that all the different name variants of the selected researcher, previously used names and names that may have been incorrectly indexed in the databases are taken into account.
- By affiliations: by defining the units whose publications constitute the data under review. During the data collection phase, care should be taken to ensure that all the current, former and multilingual spellings of the names used by the unit are taken into account.
- By search terms: the data can be collected by means of listed search terms. For example, if you wish to examine publications on sustainable development, you can start by collecting a list of key terms related to the topic.
- By publication channels: the data can be collected e.g. by means of selected scientific journals.

Regardless of how the data collection is carried out, it is advisable that those being evaluated should, as far as possible, be able to check both the data used and the results of the analysis.

Selection of data source for the analysis

To analyse publications and their citations, a citation database or analysis tool based on them is required. When making the choice, the differences between the citation databases should be considered, especially in terms of the coverage of the data they contain. It is advisable to choose the citation database that best covers the publications of the unit that is being reviewed. The quality of the data in the database, as well as the search functions and analytical features, are also important. The most used multidisciplinary citation databases and analysis tools are presented in the chapter Data sources and tools of this guide.

Further reading

A broader description of analytical processes in the context of different bibliometric reviews is presented in the following article:

Naveen D, Satish K, Debmalya M, Nitesh P, Weng ML. How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, Vol 133,2021, Pages 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Sources

Hyrkkänen, A-K. (2022) Julkaisuperusteisen analyysin suunnittelu, aineistonkeruu ja tiedonlähteiden valinta [kaavio]. Perustuu Visser M. and Calero C. CWTS Leiden kurssimateriaaliin *Measuring Science and Research Performance,* 21st September 2015.

Structure of this chapter is based on CWTS Leiden course material: Visser M. and Calero C. *Measuring Science and Research Performance* (2015) CWTS Leiden Course material, 21st September 2015