



**Karolinska
Institutet**

Bibliometrics

Publication Analysis as a Tool for Science Mapping and Research Assessment

The Karolinska Institutet Bibliometrics Project Group

Karolinska Institutet University Library

2008-10-09, version 1.3

What is bibliometrics?

Bibliometrics is the application of mathematical and statistical methods to publications (from *biblos*: book and *metron*: measurement). Bibliometrics is often used to assess scientific research through quantitative studies on research publications.

Bibliometric assessments are based on the assumption that most scientific discoveries and research results eventually are published in international scientific journals where they can be read and cited by other researchers. The number of citations to a journal article can be considered to reflect the article's impact on the scientific community.

Applied bibliometrics, as it is used today, analyzes the number of scientific articles published by a selected number of authors, citations to these articles and connections between articles, authors and subjects.

What do bibliometric analyses measure?

Bibliometric analyses result in *indicators* of research quantity and performance. They can also provide measurements of connections between researchers and research areas through statistical analysis of co-publications and citations.

This part gives a short description of which bibliometric indicators an analysis can provide. The next part will describe how to use data to produce the different indicators.

Quantity indicators: Number of publications and citations

Examples:

- **Number of publications and citations.** The two most basic bibliometric indicators describe the number of publications and citations attributed to a group of authors (a research group, a department, a university or a country) during a specified time period.
- **Number of publications and citations per researcher** is a relative measure. It compensates for the size of the studied unit and therefore indicates scientific output in relation to invested resources.

Performance indicators: Normalized citation counts

Examples:

- **The crown indicator** measures the research impact of a group of authors. It compares the average number of citations to the group's publications to the average number of citations to international publications from the same year, in the same subject area and of the same document type.
- **Top 5 %** shows the share of publications attributed to a group of authors that belong to the 5% most cited publications in the world from the same year, in the same subject area and of the same document type.

Journal Performance Indicators: Impact Indicator

- An **impact indicator** for a scientific journal is a mean value that describes how many times an average article published in the journal is cited.

Structural Indicators: Publication Patterns

Publication and citation analysis can also identify connections between publications, authors and areas of research.

One example is the use of connection maps to illustrate how much different units publish together or how a selected number of publications are connected through a common field of research.

How do you perform a bibliometric analysis?

Most bibliometric analyses use data originating from one or more of the three ISI citation indices supplied by Thomson Scientific. (ISI – the Institute for Scientific Information – founded by Eugene Garfield in 1958 and now a part of Thomson Scientific.)

The most important citation index for medicine, life science and the natural sciences is the Science Citation Index Expanded (SCIE). This contains references to articles from more than 5 900 scientific journals ("Thomson Scientific: Citation Products"). There is also a Social Sciences Citation Index and an Arts and Humanities Citation Index.

Including all three indices, Thomson Scientific indexes about 8 500 of an estimated number of more than 22 000 active, refereed scientific journals (Ulrichsweb.com).

Since Thomson Scientific use the reference lists from publications in their own indices to select what journals to include, it is reasonable to assume that the Thomson citation indices contain the most cited and most important academic journals.

Subscribers to the Thomson citation indices can, for example, access them through the web based service Web of Science. This is a relatively easy-to-use search interface that provides the opportunity to create lists of publications and citations attributed to researchers, research groups, departments, universities or countries. It's however not suited for more complex bibliometric analyses, including the calculation of mean values or connection mapping. For this you have to purchase or download data from the Thomson citation indices and use locally developed applications for the calculations.

As a consequence of the strong bibliometric focus on data from the Thomson citation indices, most bibliometric indicators are reliable only in research areas where publishing in scientific journals is the main mode of communication. This is often the case in natural sciences, technology and medicine, but analyses of areas within the humanities or social sciences must apply other methods as well.

Selecting a unit of analysis

The starting point in a bibliometric analysis is to select a group of publications, usually on the basis of information available in the Thomson citation indices. This selection of publications forms the unit of analysis.

The publications may for example be selected on the basis of the authors' organizational affiliations and can theoretically be:

- Author
- Research group
- Department
- Research Centres/Networks
- Universities

- Countries

A substantial amount of local data preparation and verification is necessary in order to create a unit of analysis based on a research group, a department, a research centre or a research network. This information is very difficult to locate in data from the Thomson citation indices and may in many cases not be present at all. It is even difficult to attribute publications to a particular university since both organization names and their addresses may be written in many different ways and two different universities occasionally share a common name.

A unit of analysis can also be selected based on the properties of individual articles (instead of authors or author affiliations).

- Individual publications
- Journal
- Subject – often based on subject classification of the journal
- Document type – article, review, note, letter, conference proceeding, etc.
- Publication year

Since statistical methods are used in bibliometric research, the results improve with larger units of analysis. This is partly because isolated phenomena – i.e. negative citations – are cancelled out by the large amount of articles (Henk F. Moed, 2005, p. 80). Bibliometric indicators based on any unit of analysis that contains less than 10 articles (i.e. an individual researcher or article) is not to be recommended ([Anonymous], 2005, p. 10).

It is also necessary to take into consideration any possibility of a systematic bias. This could for example be different citing traditions or conventions for including and ranking authors that vary significantly between different research areas. (Henk F. Moed, 2005, p. 223).

Bibliometric indicators

Many bibliometric researchers stress the importance of not considering the results from any bibliometric analysis to be "truths". The term *bibliometric indicators* is often used to note the fact that the results describe a too complex reality to be measured merely by statistics or numbers. Bibliometric methods contain so many simplifications that they only supply a very limited picture of the research they are trying to describe.

It is important to see bibliometric indicators as one of several tools to be used by competent reviewers with specific knowledge about the research areas included in the analysis. This is for example evident when publications containing very new or unconventional research results are included in an assessment. These will not yet have been cited, which means that any assessment based solely on bibliometric indicators will not discover the possible potential of the research groups in question.

No bibliometric indicator should be put to isolated use. Several indicators should always be combined to achieve a more comprehensive picture of the scientific production of a unit (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003). The Crown indicator should for example always be accompanied by a so called top indicator that shows if the mean value of citations to the unit's publications is due to a few very highly cited articles or a majority of publications cited a bit above average, and by a quantity indicator to show how many publications that are included in the analysis.

Number of publications and citations

Two very basic bibliometric indicators are the number of publications and citations during a specific time period. These two indicators do not compensate for the size of the publishing unit or the document type of the publications. However, they can be useful to someone with knowledge of the research area under study, especially if the indicators are used to compare similar research units or as a complement to other bibliometric indicators.

Number of publications and citations per researcher

Publication and citation counts in relation to the number of active researchers or employees at the studied unit are two somewhat more refined indicators of scientific production and impact.

Performance indicators based on a relative number of citations

To get more sophisticated performance indicators you need to compensate for some properties of the studied publications and compare their citation numbers to the world average.

- Publication year – citations accumulate with age which means that older articles are more highly cited.
- Document type – The number of citations to different document types varies significantly. Review articles, for example, generally receive more citations than regular articles.
- Subject – the citation patterns are different in different research areas

To compensate for these differences, an analysis resulting in performance indicators includes normalization, i.e. a comparison of publication citation counts to the citation count average of publications of the same document type, published the same year, in the same subject.

Self Citations

Self citations, citations where authors refer to their own papers, are also an aspect of bibliometric analyses focused on citation based performance indicators. Studies have shown that self citations do not significantly influence analysis results when you study a sufficiently large number of publications. This is probably because most researchers refer to their own work in equal quantity as a natural part of scientific communication. On group level however, small differences in citation counts may indeed influence indicators.

Available options to compensate for this are:

- trying to exclude self citations when calculating indicator values
- noting them so that the interpretation of the indicators can be affected by the amount of self citations
- assuming that they are evenly distributed and hence ignoring their effect when calculating the indicators

It is very difficult to remove self citations when calculating indicators and it requires data from a comprehensive citation database such as the Thomson citation indices.

The aspect of self citations has been left out in the indicator descriptions below.

The crown indicator

The crown indicator compares the average number of citations to the analyzed unit's publications to the average of citations to international publications from the same year, in the same subject and of the same document type.

It is usually written as a decimal number that shows the relation to the world average, i.e. 0.9 shows that the analyzed publications are cited 10% less than the world average and 1.2 that they are cited 20% more.

Top 5%

Top 5% shows the share of publications attributed to a group of authors that belong to the 5% most cited publications in the world from the same year, in the same subject and of the same document type.

It is just as the crown indicator written as a decimal number that shows the relation to the world average. A value over 1 shows that the analyzed unit has more of its publications among the top 5% than the world average, a value below 1 that it has less.

Top 5% is often used as a complement to the crown indicator. It shows if a high crown indicator value is achieved through a few very highly cited articles or a larger number of articles cited above average. It may also identify highly cited articles from a group with a low crown indicator value whose top publications would otherwise have been unnoticed.

Knowledge in the specific research area is required to decide which of the two publication patterns is the better sign of high-quality research.

Uncited publications

The share of publications that remain uncited after a certain time period can be considered the opposite of the top 5% indicator. If the crown indicator of a group's publications has a high value, information about a large number of uncited publications implies that most effort has been put into a few "flag papers".

ISI Journal Impact Factor

The ISI impact factor was designed by Eugene Garfield around 1960 as a means to measure the impact of a specific journal, and it gives an average value on how many times an article in the journal has been cited. It is defined as the average number of citations given in a specific year to documents published in that journal in the two preceding years, divided by the number of documents published in that journal in those two years.

It is used by ISI to localize the most important scientific journals in each research area and is in this respect also used as a library collections management tool.

The impact factor is often written as a number corresponding to the average number of citations to an article in the journal (including the document types reviews and notes). The journals with the highest impact factors can reach numbers of around 50, and journals like New England Journal of Medicine, Nature and Science have impact factors of about 30. Most journals have an impact factor below 1.

The fact that the ISI Impact factor is based on citations only 1-2 years old can be considered a compromise between the need of getting a quick appraisal of new journals and letting the publications reach their citation maximum (the year when the publication receives most of its citations). Most articles reach their citation maximum 3-5 years after publication, so that would for many research areas be a preferable citation window.

The citation patterns vary so much between different research areas that the ISI impact factor should not be used to compare the scientific impact of journals in different subjects.

Since the ISI impact factor is relatively easy to find and understand it has become very popular and is often used to assess the quality of articles, researchers, departments and universities by studying the journals they publish in. This is inadvisable and not what the impact factor was intended for.

However, if you wish to study very recently published articles that have not yet been cited, a journal impact indicator may be the only possible performance indicator. The indicator is then based on the assumption that the refereeing process is more rigorous in high-impact journals, which means that only high-quality research will be accepted.

Number of publications in high-impact journals

Publications in high-impact journals are often considered to be of high quality. It is not uncommon for researchers to be asked to supply information about their "mean impact factor", i.e. the average value of the impact factors of the journals they have published in, when they apply for a grant or a new position.

Sometimes a research unit or a university also displays how many publications they have in journals with very high ISI impact factors, i.e. the 20 or 40 most highly ranked, as a sign of quality research produced by authors at that unit. This measurement is usually not compensated for the size of the analysed unit, which means that larger institutions get higher values. The research areas of the analysed unit will also affect this figure, so sometimes different journal impact factor lists are used for different research areas.

However, the impact factor of a journal cannot predict the number of citations that any individual publication will receive. Often about 20% of the publications in a journal receive 80% of the citations and many articles are cited 0-1 times even in high impact journals.

Publication patterns

Co-publication

If the publication data used for analysis contains all author address (as in data from the Thomson citation indices) it can be used to find patterns of co-publication between different authors and units. These can be visualized in co-publication tables and maps that show for example the authors, universities and countries that publish together and to what extent.

Subject interrelatedness

If data contains a subject classification of the publications, statistical methods can be used to analyse for example:

- Which subjects a researcher specifies in and how these connect him to other authors.
- Research areas that are connected by a certain number of publications.
- If research areas that show up frequently in the publications of a department are organized in one large departmental research network or several smaller ones.

Next generation indicators

Current bibliometric indicators have made it possible to relate the number of citations received by a publication to a world average based on subject area, time since publication and document type. This has been a major step towards an increased validity and applicability of bibliometric indicators. However, it should be noted that these indicators are not final, and new indicators are under development. Examples of indicator development currently performed at Karolinska Institutet are:

- New subject classification: Current indicators use a subject classification based on journal subject categories supplied by Thomson. We have previously shown that this subject classification has several shortcomings (J. Lundberg, Fransson, Brommels, Skar, & Lundkvist, 2005). Therefore Karolinska Institutet is testing the possibility to use MeSH-terms as the basis for subject classification on an article level (Jonas Lundberg, 2006).
- New statistical method: To calculate the state-of-the-art indicator “crown” the received number of citations is divided by the expected number (the world average for publications of the same type, from the same year, within the same area). This does not take the citation distribution of different areas into account. Therefore Karolinska Institutet is exploring the possibility to use other statistical methods, for example z-score (Jonas Lundberg, 2006).

How is bibliometrics used at Karolinska Institutet?

The possible use of data from Thomson ISI to analyse research publications from Karolinska Institutet has been studied since 2002. In 2005 a bibliometric pilot study was performed in cooperation with the Swedish Research Council.

In late 2005 the management of Karolinska Institutet decided that bibliometrics will be used as a tool in the ambition to become the leading medical university in Europe by 2010. This led to the formation of the project Karolinska Institutet Bibliometrics in the beginning of 2006.

The project Karolinska Institutet Bibliometrics

The project uses data about international scientific publications from the Thomson citation indices 1995-2008 to build a system capable of analysing Karolinska Institutet research publications and compare these to international publications. In the beginning of 2006, 10 million publication records from 1995-2005 were bought and loaded into a database. Additional records are downloaded weekly into the database as they are delivered from Thomson Scientific and the current Karolinska Institutet bibliometric database is expected to be as updated as the original Thomson databases.

Karolinska Institutet researchers are regularly asked to log in to an internal web site and verify their publications in the bibliometric database, something that increases the quality of both data and bibliometric analyses of Karolinska Institutet publications.

The primary goal is to supply the Karolinska Institutet management with analyses on Karolinska Institutet research publications as part of the institute's quality management. Secondary goals are to give researchers and employees working at the institute the opportunity to order analyses and to spread information about Karolinska Institutet publications and publication patterns.

A bibliometry-oriented customer service

Apart from analyses made for the Karolinska Institutet management, researchers and other employees of the institute have the possibility to order bibliometric analyses from the University Library at an individual cost.

A Karolinska Institutet bibliometric handbook

To our knowledge, bibliometrics has not before been used to continuously assess an organization in the way that is now put in production at Karolinska Institutet. Advanced bibliometric research has been performed at departments with research in sociology, library and information science and statistics, but since practical applications are rarely made, indicators have not been standardized, but rather continuously developed. This is one of the reasons why the project group at Karolinska Institutet is compiling a "Bibliometric Handbook for Karolinska Institutet" where the group describes how the different indicators used by Karolinska Institutet are produced. The handbook also describes the advantages and disadvantages of different indicators and methods and give instructions on how to interpret the indicators and other results of bibliometric analysis. The handbook and other publications from the Karolinska Institutet Bibliometric Project can be downloaded from the external project site at <http://ki.se/bibliometrics>.

Additional recommended reading

New Bibliometric Tools for the Assessment of National Research Performance - Database Description, Overview of Indicators and First Applications (H. F. Moed, Debruin, & Vanleeuwen, 1995). An early description of the database and many of the bibliometric indicators developed at CWTS, Leiden.

Bibliometrics as a research field: A course on theory and application of bibliometric indicators. (Glänzel, 2003)

Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003; Van Raan, 2005).

References

- [Anonymous]. (2005). *Quantitative Indicators for Research Assessment - A Literature Review* (Review No. 05/1). Canberra: Research Evaluation and Policy Project, Research School of Social Sciences, Australian National University
- Glänzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators.
- Lundberg, J. (2006). Lifting the Crown. *Forthcoming*.
- Lundberg, J., Fransson, A., Brommels, M., Skar, J., & Lundkvist, I. (2005). Is it better or just the same? Article identification strategies impact bibliometric assessments. *Scientometrics*, 66(1), 183-197.
- Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.
- Moed, H. F., Debruin, R. E., & Vanleeuwen, T. N. (1995). New Bibliometric Tools for the Assessment of National Research Performance - Database Description, Overview of Indicators and First Applications. *Scientometrics*, 33(3), 381-422.
- Thomson Scientific: Citation Products. Retrieved April 6th, 2006, from <http://scientific.thomson.com/products/categories/citation/>
- Ulrichsweb.com. Advanced Search. Retrieved April 6th, 2006, from <http://www.ulrichsweb.com/ulrichsweb/Search/advancedSearch.asp?>
- van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & van Raan, A. F. J. (2003). Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, 57(2), 257-280.
- Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133-143.