

# Bibliometric analysis on repository contents as a library service for the evaluation of research



**Marco G.P. Van Veller, Wouter Gerritsma**  
Wageningen UR Library, The Netherlands  
marco.vanveller@wur.nl  
gerritsma@wur.nl

## Abstract

Since the last two decennia, the library of Wageningen University and Research (or Wageningen UR) has been involved in several bibliometric analysis for the evaluation of scientific output of staff, chair groups, research institutes and graduate schools. In these bibliometric analysis several indicator scores, such as the number of publications, number of citations and citation impacts, are calculated. However, since publication and citation habits differ considerably per scientific discipline, indicator scores need to be standardized in order to make comparison of scientific outputs from staff, chair groups, research institutes or graduate schools, that belong to different scientific disciplines, possible. It is for this reason that Wageningen UR Library bases its bibliometric analysis on indicator scores that are standardized against average trend (or baseline) scores per research field. For the collection of scientific output that is subjected to bibliometric analysis the repository Wageningen Yield (WaY) is used. This repository contains metadata and links to scientific output of the staff, research groups, chair groups and institutes of Wageningen UR and is developed and managed by Wageningen UR Library. By the application of a link between the metadata of publications in WaY and citation scores in Thomson Reuters' Web of Science, custom-made analysis can be performed efficiently. Using the repository WaY as a source for Wageningen UR's scientific output enables optimal use of meta data (e.g. affiliations of authors), improvement of completeness and data quality as well as awareness of library services. Moreover, the bibliometric analysis give insights in steps that can be taken to improve publication strategies for Wageningen UR staff, chair groups, research institutes and graduate schools.

### **Keywords**

Bibliometric analysis, Citation scores, Repositories

### **Résumé**

#### **Analyse bibliométrique des contenus d'archives ouvertes : un service de la bibliothèque pour l'évaluation de la recherche**

Depuis vingt ans, la bibliothèque de l'Université de Wageningen (Pays-Bas) a contribué à plusieurs analyses bibliométriques pour l'évaluation de la production scientifique de chercheurs, équipes, instituts de recherches et facultés. À cet effet, plusieurs mesures ont été utilisées (nombre de publications, nombre de citations, impact des citations). Cependant, en regard des différences de pratiques de publication qui s'inscrivent dans la discipline, les indicateurs doivent faire l'objet d'une normalisation qui rend possible la comparaison de la production scientifique entre disciplines. C'est pourquoi la BU de Wageningen fonde ses analyses bibliométriques sur des indicateurs pondérés par la tendance moyenne du domaine de recherche. Le dépôt institutionnel de l'université, Wageningen Yield (WaY), recueille les publications évaluées des chercheurs. Ce dépôt qui contient les métadonnées de ces publications et les liens vers leur texte intégral est géré par la bibliothèque. En reliant les métadonnées des publications disponibles dans l'archive ouverte de l'Université avec les citations du Web of Science, des analyses pointues peuvent être réalisées. L'utilisation du dépôt WaY en tant qu'archive et dépôt des travaux publiés par l'université permet une exploitation optimale des métadonnées (notamment l'affiliation des auteurs), soulignant ainsi de souligner la valeur ajoutée des services de bibliothèques. Mais surtout, les analyses bibliométriques contribuent à dessiner les stratégies de publication de la recherche à l'université.

### **Mots-clés**

Citations, archives ouvertes, bibliothèque

Marco VAN VELLER et Wouter GERRITSMA sont bibliothécaires spécialisés en ressources électroniques à la bibliothèque universitaire de Wageningen, aux Pays-Bas.

One of the main activities of doing scientific research involves the publication on this research in scientific literature. It is for this reason that measurement of scientific output and impact of scientific publications may be used to estimate the productiveness and importance of the underlying research activities. Both productiveness and importance of research are essential criteria for making decisions regarding further development of research at institutes or universities. Therefore, policy makers and boards of these research organizations are interested in quantitative measurements on the output and impact of scientific publications for monitoring and evaluation purposes.

Since the last two decennia, the library of Wageningen University and Research (or Wageningen UR) has been involved in various quantitative analysis on scientific publications produced by staff of the university and institutes. These advanced bibliometric analysis were performed for various periods of time on scientific output of staff, chair groups and research institutes (or entities) of Wageningen UR.

Originally, the scientific output, on which the bibliometric analysis were performed, was selected by doing a retrospective search in citation enhanced bibliographical databases for the Wageningen UR entities under consideration. Thereby it was very important to be as complete as possible in the retrospective search in order to select all scientific output on which the analysis should be based. For this reason, the results from the retrospective search were checked by the Wageningen UR entities for which the bibliometric analysis was performed. Missing publications were added and corrections were made before the scientific output was subjected to the bibliometric analysis.

The additions and corrections on scientific output, however, were not registered and needed to be made for each subsequent bibliometric analysis. It is for this reason that since 2009 the library of Wageningen UR uses the repository Wageningen Yield (WaY) for bibliometric analysis for Wageningen UR entities. This repository contains the meta data of all publications (as well as a link to the full-text of the publication if available) that have been published by the staff and employees from Wageningen UR.

For this reason WaY is a very comprehensive source for the scientific output that is subjected to the bibliometric analysis. This paper discusses the application of the repository in the bibliometric analysis performed by Wageningen UR Library as well as the mutual advantages to both the bibliometric analysis and the repository that results from this combination.

## ◆ 1. Choice of bibliographic databases for bibliometric analysis

The aim of a bibliometric analysis is to obtain a quantitative overview of the scientific output and the impact of citations to this scientific output published by a researcher, research group or research organization. At Wageningen UR bibliometric analysis are made at regular intervals as part of the continuous research evaluation of staff, research groups or institutes. The advanced bibliometric analysis performed by Wageningen UR Library are based upon scientific articles that are published in scholarly journals or serials that are covered by the bibliographic database Web of Science (WoS) from Thomson Reuter Scientific.

In WoS almost 8000 current peer reviewed journals in different scientific disciplines are covered. This selection of journals is considered to include the most important peer reviewed journals for science. In addition to standard bibliographical data, WoS contains for each of the articles (in the covered journals) the number of citations to each article. The latter is updated on a regular basis.

Besides WoS, Thomson Reuters Scientific also produces the database Essential Science Indicators (ESI) which provides the world average number of citations of articles for 22 different research fields. The world averages are updated five times per year and a series of world averages over a certain number of years is called a baseline. Besides world averages, ESI also provides the average number of citations per year for the top 0.01 %, 0.1 %, 1 %, 10 %, 20 % and 50 % most cited articles per research field. In an advanced bibliometric analysis combination of WoS and ESI is essential for the calculation of the impact of citations because citation counts for an article should be obtained from WoS, the baselines follow from ESI and the journals are classified in the research fields in ESI.

## ◆ 2. Normalization of citation impacts

The impact of citations to an article is normalized for one of the 22 different research fields to which the article belongs. Thereby, the research field for an article is determined on basis of the journal in which it has been published and can be found in the ESI. For each article, the citation impact is normalized by benchmarking the number of citations against the world average citation value, obtained for the same year in which the article has been published via the baseline for the research field to which the article belongs. The reason to normalize citation impacts lies in the different publication and cita-

tion cultures that apply to different research fields (*e.g.* the world average number of citations for publications in Molecular biology & Genetics are approximately eight times larger than in Mathematics). It is for this reason that only by normalizing citation impacts, monitoring and evaluation of these impacts over different research fields can be performed.

Normalization of citation impacts of a series of articles (*e.g.* published by a particular research group that is subject to a bibliometric analysis) can be done in two alternative ways. Both ways of normalization result in a measure that shows the relation of the impact of the analyzed set of articles to the world average of 1. One normalized measure for the citation impact is obtained by dividing the sum of citations to all considered articles by the sum of the world average citation rates for these articles. This ratio leads to the Citation Impact (CI or Crown Indicator sensu VAN RAAN, 2004) for the series of articles. Alternatively, normalization of citation impacts of a series of articles can also be carried out at the level of individual articles. In this case, a citation impact is calculated for each article separately and normalization is accomplished by calculation of the mean of these impacts. This approach on the level of individual articles leads to a measurement for the citation impact known as the Relative Impact (RI sensu GERRITSMA, 2006).

### ◆ **3. Advanced indicators for bibliometric analysis**

Besides normalized citation impacts also other indicators result from a bibliometric analysis. In bibliometric analysis at Wageningen UR Library for a series of articles (*e.g.* for a particular Wageningen UR entity) the following indicators are measured:

- N = Total number of articles in series that is examined =  $\sum_{i=1}^T n_i$

- C = Total number of citations to the N articles =  $\sum_{i=1}^T \sum_{j=1}^{n_i} C_{t,j}$

- CPP = Average number of citations per article =  $\frac{\sum_{i=1}^T \sum_{j=1}^{n_i} C_{t,j}}{N}$

- CI = Citation Impact =  $\frac{\sum_{i=1}^T \sum_{j=1}^{n_i} C_{t,j}}{\sum_{i=1}^T \sum_{j=1}^{n_i} Wavg_{t,j}}$

- RI = Relative Impact =  $\frac{\sum_{i=1}^T \sum_{j=1}^{n_i} \frac{C_{t,j}}{Wavg_{t,j}}}{N}$

- %T10 = Percentage of the 10 % most cited articles compared to the total number of articles in the series =  $\frac{N(C_{t,i} \geq C_{10\%,t,i})}{N} * 100\%$
- %T1 = Percentage of the 1 % most cited articles compared to the total number of articles in the series =  $\frac{N(C_{t,i} \geq C_{1\%,t,i})}{N} * 100\%$
- %NC = Percentage of non-cited articles compared to the total number of articles in the series =  $\frac{N(C_{t,i} = 0)}{N} * 100\%$

With:

$n_t$  = Number of articles published in year t

$C_{t,i}$  = Number of citations to article i, published in year t

$Wavg_{t,i}$  = World average number of citations for articles in same research field as for article i for the year t in which article i has been published

$C_{10\%,t,i}$  = Average number of citations for 10 % best cited articles in same research field as for article i for the year t in which article i has been published

$C_{1\%,t,i}$  = Average number of citations for 1 % best cited articles in same research field as for article i for the year t in which article i has been published

Via these eight indicators Wageningen UR Library tries to give an impression of both the scientific productivity (by counting the number of articles published by the Wageningen UR entities) as well as the impact of this output in the form of citations (by calculation of Citation or Relative Impacts of articles published by the Wageningen UR entities).

For the eight advanced bibliometric indicators both the CI and RI are preferred for the measurement of citation impact because they have been normalized to the research field in which the Wageningen UR entity is working. Because of the different way of normalizing citation impacts of articles differences can be found between these two indicators. Hereby, the RI measures citation impact on an individual basis and may be susceptible to outliers due to articles with high citation scores related to their corresponding world averages. The CI, to the contrary, applies normalization to the complete set of articles as a whole and thereby gives a relative larger weight to the number of citations for older articles and articles in highly cited research fields.

Although the CPP is not normalized to a particular research field, it gives (in relation to the CI) an impression on the world average number of citations per article for the most dominant research field in which the Wageningen UR entity has published. In general, the CPP shows an increasing trend with the age of the articles for which the CPP has been calculated. Both the %T10 and %T1 give indications for excellent cited Wageningen UR entities. Finally,

high values of %NC weigh down the values of CPP, RI and CI in bibliometric analysis.

#### ◆ **4. Repository contents as a source for bibliometric analysis**

In order to perform a bibliometric analysis on the scientific output of a particular Wageningen UR entity, first the series of articles for which the eight different indicators will be calculated need to be selected. As indicated above, these articles need to be published in scientific journals or serials that are covered by WoS. By examining these articles only, one is able to obtain the number of citations for each article from WoS and relate it to values of the applicable baseline for normalization.

Besides selection of articles from journals covered by WoS it is important that the series of articles that has been selected is identified as the set of articles published by the Wageningen UR entity that is evaluated over a certain period of time. Hereby it is important that a complete set of articles is obtained and that no mistakes are made by selection of articles from other persons, groups or institutes than the one(s) that is evaluated. Selection of the right series of articles for a particular Wageningen UR entity is based upon the meta data (such as affiliation information of the authors) of the articles.

Meta data of the scientific output of whole Wageningen UR are collected in a research registration system (called Metis) and made accessible to users *via* the repository WaY. Registration of new publications in the research registration system is done by the secretaries of the research groups and both the research registration system Metis and the repository WaY are (further) developed and maintained by Wageningen UR Library.

The repository WaY contains besides the meta data also links to scientific output of the Wageningen UR entities. Further, also affiliation data and all work relations of the staff of Wageningen UR with various Wageningen UR entities (and others outside Wageningen UR) and the research theme or graduate school in which the staff member participates is recorded in great detail. The in-house registration and dissemination of meta data of scientific output from Wageningen UR staff enables a more reliable selection of a series of articles for a bibliometric analysis (for a particular Wageningen UR entity) than selection of these articles on the basis of affiliations mentioned in the articles. Also, if the registration of meta data for new scientific output in the repository WaY is complete and timely, the chances of missing articles in a bibliometric analysis are minimized.

## ◆ **5. Connection between the repository and bibliographical database for bibliometric analysis**

For a bibliometric analysis of a series of articles (published by a particular Wageningen UR entity) selected from the repository WaY a connection needs to be made between the meta data for the selected articles in the research registration system Metis and the bibliographical information for the same articles in WoS.

By searching in WoS on affiliation Wageningen UR and its various institutes a list of articles from authors that work (or have worked) at Wageningen UR is obtained. The bibliographical records for these articles are matched with records for the articles that are registered in Metis by using information from the Digital Object Identifier, ISSN in combination with volume and page number or title.

As a result of this matching for each article in Metis (which can be identified in Metis by a unique article number), its ISI-number is obtained from WoS. The ISI-number is entered in the record for the article in WaY and can be used for identification of the article in WoS. Since the ISI-number of each article is stored in WaY the matching between bibliographical data from Metis with WoS needs to be done for each (added) article only once.

With the meta data in WaY on affiliation and work relations different selections can be made for series of articles published by one (or more) Wageningen UR entity (or entities) that will be subjected to a bibliometric analysis. For a particular selected series of articles ISI-numbers are collected from WaY and used to obtain the number of citations ( $C_{t,i}$  in formulas above) per article from WoS.

Because this number of citations changes over time it needs to be collected from WoS on a frequent basis. Besides the number of citations also the world average number of citations ( $Wavg_{t,i}$  in formulas above) and the average number of citations for 10 % and 1 % most cited articles (respectively  $C_{10\%,t,i}$  and  $C_{1\%,t,i}$  in formulas above) are updated for each research field from ESI frequently.

The values of the bibliometric data  $C_{t,i}$ ,  $Wavg_{t,i}$ ,  $C_{10\%,t,i}$  and  $C_{1\%,t,i}$  are used to calculate the different indicators that are mentioned in the formulas above. Hereby the research field of an article is obtained by making a connection (on basis of ISSN; obtained from WaY) between the journal in which the article has been published and information obtained from ESI on the classification of journals in the different research fields.



## ◆ 6. Example of a bibliometric analysis for a research group of Wageningen UR

An example of the results from a bibliometric analysis that has been performed for the scientific output of a research group of Wageningen UR is given in the two tables below. Table 1 gives yearly results for the different indicators whereby publications from different research field may have been combined per year. Table 2 is based on the same series of articles from this research group but lists the different indicators per research field whereby publications from different years may have been combined in a research field. From the two tables trends in the scientific output and impact over years can be derived as well as information on the most prominent research fields in which this research group published.

**TABLE 1 – INDICATORS CALCULATED FROM A BIBLIOMETRIC ANALYSIS ON 139 ARTICLES PUBLISHED BY A RESEARCH GROUP FROM WAGENINGEN UR IN THE YEARS 2002-2007**

Year	N	C	CPP	CI	RI	%T10	%T1	%NC
2002	10	527	52.7	3.46	4.92	50 %	30 %	0 %
2003	21	545	26.0	1.54	1.89	33 %	5 %	0 %
2004	25	509	20.4	1.71	2.53	28 %	8 %	0 %
2005	25	323	12.9	1.35	1.83	28 %	4 %	0 %
2006	26	319	12.3	2.52	3.22	38 %	15 %	8 %
2007	32	285	8.9	3.73	4.69	38 %	19 %	13 %
All years	139	2508	18.0	2.01	3.10	35 %	12 %	4 %

**TABLE 2 – INDICATORS CALCULATED FROM A BIBLIOMETRIC ANALYSIS ON 139 ARTICLES PUBLISHED BY A RESEARCH GROUP FROM WAGENINGEN UR IN FIVE DIFFERENT RESEARCH FIELDS**

Year Research Field	N	C	CPP	CI	RI	%T10	%T1	%NC
Plant & Animal Science	73	1427	19.6	4.08	4.29	49 %	19 %	4 %
Molecular Biology & Genetics	38	686	18.1	1.03	1.17	18 %	0 %	0 %
Biology & Biochemistry	21	365	17.4	1.85	2.69	19 %	10 %	14 %
Environment/Écology	6	16	2.7	0.54	0.78	0 %	0 %	0 %
Agricultural Sciences	1	14	14.0	12.96	12.96	100	100	0 %
All research fields	139	2508	18.0	2.01	3.10	35 %	12 %	4 %

## ◆ **7. Mutual advantages to the use of repository contents in bibliometric analysis for Wageningen UR entities**

The application of the Wageningen UR repository WaY for the selection of articles for the evaluation of scientific output and citation impact of Wageningen UR entities gives advantages to the efficiency and quality of both the repository and the bibliometric analysis.

For the repository, complete coverage and timely registration of new publications is stimulated because by registration of these publications in WaY, the Wageningen UR entities ensure that also their new scientific output will be included in future bibliometric analysis. The meta data in WaY is entered by the secretaries and checked by the research staff of Wageningen UR at the registration of new publications in Metis as well as double checked on a regular basis by staff from Wageningen UR Library. Besides these checks, errors in the meta data and coverage of the scientific output from Wageningen UR in WaY is checked when bibliographical records in Metis and WaY are matched with the corresponding ones in WoS. Additions as well as corrections that are made on the bibliographic records for Wageningen UR scientific output are registered in Metis and WaY and thereby are ensured for inclusion in future analysis and uses of meta data such as generation of publication lists, etc.

For the advanced bibliometric analysis that Wageningen UR Library performs on Wageningen UR entities, quality is improved by a more precise selection of the right articles for the entities under evaluation due to reliable and updated registration of affiliation data and work relations. Moreover, by using the repository for selection of scientific output, custom-made bibliometric analysis can be performed almost “on the fly” and comparative research evaluations can be performed in a fast and efficient way.

By the establishment of a connection between meta data on affiliation and work relations in the repository WaY and bibliometric data from WoS and ESI a system has been developed in which bibliographical information on new publications is entered when they are registered in the research registration system Metis. For this system updates on the bibliometric data can be performed efficiently and on a regular basis. Finally, the bibliometric analysis performed by Wageningen UR Library in this way has improved awareness of library services considerably. Wageningen UR Library is consulted for questions with regard to steps that can be taken to improve publication strategies for staff members, research groups or institutes.

## ◆ **References**

GERRITSMAN, W., (2006). « Een systeem voor citatieanalyses in de praktijk ». *InformatieProfessional*, 10, 12-17

VAN RAAN, A. F. J., (2004). « Measuring Science. Capita Selecta of Current Main Issues ». *In*: Moed, H.F., Glänzel, W., and Schmoch, U. (eds). *Handbook of Quantitative Science and Technology Research: The use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht: Kluwer Academic Publishers. P. 19-50.