CWTS Meaningful metrics

# CWTS Leiden Ranking 2015

## Methodology

Universiteit Leiden

Centre for Science and
Technology Studies,
Leiden University

# Data collection

The CWTS Leiden Ranking 2015 ranks the 750 universities worldwide with the largest publication output in international scientific journals in the period 2010–2013. The ranking is based on data from the Web of Science database. A sophisticated data collection methodology is employed to assign publications to universities.

## Web of Science

The Leiden Ranking is based exclusively on bibliographic data from the Web of Science database produced by Thomson Reuters. The ranking uses data from the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The Leiden Ranking is based on Web of Science data because Web of Science offers a good coverage of the international scientific literature and generally provides high quality data.

The Leiden Ranking does not take into account conference proceedings publications and book publications. This is an important limitation in certain research fields, especially in computer science, engineering, and the social sciences and humanities.

## Enriched data

CWTS enriches Web of Science data in a number of ways. First of all, CWTS performs its own citation matching (i.e., matching of cited references to the publications they refer to). Furthermore, in order to calculate the more advanced collaboration indicators included in the Leiden Ranking, CWTS performs geocoding of the addresses listed in publications in Web of Science and CWTS identifies addresses belonging to the business sector. Most importantly, CWTS puts a lot of effort in assigning publications to universities in a consistent and accurate way. This is by no means a trivial issue. Universities may be referred to using many different name variants, and the definition and delimitation of universities is not always obvious. The methodology employed in the Leiden Ranking to assign publications to universities is discussed in detail below.

## Identification of universities

The criteria that define universities are not internationally set, thus presenting a challenge in identifying them. Typically, a university is characterized by a combination of education and research tasks in conjunction with a doctorate-

granting authority. However, these characteristics do not mean that the universities are particularly homogeneous entities that allow for international comparison on every aspect. As a result of its focus on scientific research, the Leiden Ranking presents a list of institutions that have a high degree of research intensity in common. Nevertheless, the ranking scores for each institution should be evaluated in the context of its particular mission and responsibilities which are strongly linked to national and regional academic systems. Academic systems - and the role of universities therein - differ substantially from one another and are constantly changing. Inevitably, the outcomes of the Leiden Ranking reflect these differences and changes.

The international variety in the organization of academic systems also poses difficulties in terms of identifying the proper unit of analysis. In many countries, there are collegiate universities, university systems, or federal universities. Instead of applying formal criteria, when possible we followed common practice based on the way these institutions are perceived locally. Consequently, we treated the University of Cambridge and the University of Oxford as entities but in the case of the University of London, we distinguished between the constituent colleges. For the United States, university systems (e.g. the University of California) were split up into separate universities. The higher education sector in France, like in many other countries, has gone through several reorganizations in recent years. Many French institutions of higher education have been grouped together in *Pôles de Recherche et d'Enseignement Supérieur (PRES)* –and the more recent C*ommunautés d'Universités et Etablissements (COMUEs)* –or in consortia. In most cases, the Leiden Ranking still distinguishes between the different constituent institutions but in particular cases of very tight integration, consortia were treated as if they were a single university (e.g. Grenoble INP).

Publications are assigned to universities based on their most recent configuration. Changes in the organizational structures of universities up to 2014 have been taken into account. For example, in the Leiden Ranking 2015, the University of Bordeaux encompasses all publications previously assigned to the University of Bordeaux I, University of Bordeaux Segalen II, and Montesquieu University Bordeaux IV.

## Affiliated institutions

A key challenge in the compilation of a university ranking is the handling of publications originating from research institutes and hospitals associated with

universities. Among academic systems a wide variety exists in the types of relations maintained by universities with these affiliated institutions. Usually, these relationships are shaped by local regulations and practices affecting the comparability of universities on a global scale. As there is no easy solution for this issue, it is important that producers of university rankings employ a transparent methodology in their treatment of affiliated institutions.

CWTS distinguishes three different types of affiliated institutions:

1. component
2. joint research facility or organization
3. associated organization

In the case of *components* the affiliated institution is actually part of a university or so tightly integrated with it or with one of its faculties that the two can be considered as a single entity. The University Medical Centres in the Netherlands which combine the medical faculties and the university hospitals are examples of components. In these cases, all teaching and research tasks in the field of medicine–traditionally the responsibility of the universities–have been delegated to these medical centres.

*Joint research facilities or organizations* are the same as components except for the fact that they are administered by more than one organization. The Brighton & Sussex Medical School (the joint medical faculty of the University of Brighton and the University of Sussex) and Charité (the medical school for both the Humboldt University and Freie Universität Berlin) are examples of this type of affiliated institution.

The third type of affiliated institution is the *associated organization* which is more loosely connected to a university. This organization is an autonomous institution that collaborates with one or more universities based on a joint purpose but at the same time has separate missions and tasks. In many countries, hospitals that operate as teaching or university hospitals fall into this category. Massachusetts General Hospital, one of the teaching hospitals of Harvard Medical School, is an example of an associated organization.

The treatment of university hospitals is of substantial consequence given that medical research has a strong presence in the Web of Science. The importance of associated organizations is growing as universities present themselves more and more frequently as network organizations. As a result, researchers formally employed by the university but working at associated organizations may not always

mention the university in publications. On the other hand, as universities become increasingly aware of the significance of their visibility in research publications, they actively exert pressure on researchers to mention their affiliation with the university in their publications.

In the Leiden Ranking 2015, publications from affiliated institutions of the first two types are considered as output from the university. A different procedure has been followed for publications from associated organizations. A distinction is made between publications from associated organizations that also mention the university and publications from associated organizations that do not contain such a university affiliation. In the latter case, publications are not counted as publications originating from the university. In the event that a publication contains affiliations from a particular university as well as affiliations from its associated organization(s), both types of affiliations are credited to the contribution of that particular university to the publication in the fractional counting method.

## Selection of universities included in the ranking

The 750 universities included in the Leiden Ranking 2015 were selected based on their publication output in the period 2010–2013. Only so-called core publications were counted, which are publications in international scientific journals. Also, only research articles and review articles were taken into account. Other types of publications were not considered. Furthermore, collaborative publications were counted fractionally. For instance, if a publication includes three addresses of which two belong to a particular university, the publication was counted with a weight of 2 / 3 = 0.67 for that university. About 1100 fractionally counted publications were required for a university to be included in the Leiden Ranking 2015.

It is important to note that universities do not need to apply to be included in the Leiden Ranking. The universities included in the Leiden Ranking are selected by CWTS according to the procedure described above. Universities do not need to provide any input themselves.

## Data quality

The assignment of publications to universities is not free of errors, and it is important to emphasize that in general universities do not verify and approve the results of the Leiden Ranking data collection methodology. Two types of errors are possible. On the one hand, there may be false positives, which are publications that

have been assigned to a university when in fact they do not belong to the university. On the other hand, there may be false negatives, which are publications that have not been assigned to a university when in fact they do belong to the university. The data collection methodology of the Leiden Ranking can be expected to yield substantially more false negatives than false positives. In practice, it turns out to be infeasible to manually check all addresses occurring in Web of Science. Because of this, many of the 5% least frequently occurring addresses in Web of Science have not been manually checked. This can be considered a reasonable upper bound for errors, since most likely the majority of these addresses do not belong to universities.

# Main fields

The CWTS Leiden Ranking 2015 provides statistics not only at the level of science as a whole but also at the level of the following five main fields of science:

1. Biomedical and health sciences
2. Life and earth sciences
3. Mathematics and computer science
4. Physical sciences and engineering
5. Social sciences and humanities

## Algorithmically defined fields

Each publication of a university belongs to one, or sometimes to more than one, of the above main fields. If a publication belongs to more than one main field, the publication is assigned fractionally to each of the main fields. For instance, a publication belonging to two main fields is assigned to each of the two fields with a weight of 1 / 2 = 0.5.

Publications are assigned to the five main fields using an algorithmic approach. Traditionally, fields of science are defined by sets of related journals. This approach is problematic especially in the case of multidisciplinary journals such as *Nature*, *PLoS ONE*, *PNAS*, and *Science*, which do not belong to one specific scientific field. The five main fields listed above are defined at the level of individual publications rather than at the journal level. In this way, publications in multidisciplinary journals can be properly assigned to a field.

Publications are assigned to main fields in the following three steps:

1. We start with 3822 micro-level fields of science. These fields are constructed algorithmically. Using a computer algorithm, each publication in Web of Science is assigned to one of the 3822 fields. This is done based on a large-scale analysis of hundreds of millions of citation relations between publications. It should be noted that the 3822 micro-level fields play an important role in the calculation of the field-normalized impact indicators in the Leiden Ranking.
2. We then determine for each of the 3822 micro-level fields the overlap with each of the 249 journal subject categories defined in Web of Science (excluding the *Multidisciplinary Sciences* subject category).

3. Each subject category in Web of Science has been linked to one of the five main fields. Based on the link between subject categories and main fields, we assign each of the 3822 micro-level fields to one or more of the five main fields. A micro-level field is assigned to a main field if at least 25% of the publications in the micro-level field belong to subject categories linked to the main field.

After the above steps have been taken, each publication in Web of Science has an assignment to a micro-level field, and each micro-level field in turn has an assignment to at least one main field. Combining these results, we obtain for each publication an assignment to one or more main fields.

# Indicators

The CWTS Leiden Ranking 2015 is based on publications in Thomson Reuters' Web of Science database (Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index) in the period 2010–2013. Book publications, publications in conference proceedings, and publications in journals not indexed in Web of Science are not included. Within Web of Science, only so-called core publications are included, which are publications in international scientific journals. In addition, only publications of the Web of Science document types *article* and *review* are considered.

## Size-dependent vs. size-independent indicators

Except for the publication output indicator P, all indicators included in the Leiden Ranking have two variants: A size-dependent and a size-independent variant. Size-dependent indicators are obtained by counting the absolute number of publications of a university that have a certain property, while size-independent indicators are obtained by calculating the proportion of the publications of a university with a certain property. For instance, the number of highly cited publications of a university and the number of publications of a university co-authored with other organizations are size-dependent indicators. The proportion of the publications of a university that are highly cited and the proportion of a university's publications co-authored with other organizations are size-independent indicators. In the case of size-dependent indicators, universities with a larger publication output tend to perform better than universities with a smaller publication output. Size-independent indicators have been corrected for the size of the publication output of a university. So when size-independent indicators are used, both larger and smaller universities may perform well.

## Impact indicators

The Leiden Ranking offers the following indicators of the scientific impact of a university:

- *P(top 1%)* and *PP(top 1%)*. The number and the proportion of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 1% most frequently cited.
- *P(top 10%)* and *PP(top 10%)*. The number and the proportion of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 10% most frequently cited.
- *P(top 50%)* and *PP(top 50%)*. The number and the proportion of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 50% most frequently cited.
- *TCS* and *MCS*. The total and the average number of citations of the publications of a university.
- *TNCS* and *MNCS*. The total and the average number of citations of the publications of a university, normalized for field and publication year. An MNCS value of two for instance means that the publications of a university have been cited twice above the average of their field and publication year.

Citations are counted until the end of 2014 in the calculation of the above indicators. Author self citations are excluded. All indicators except for TCS and MCS are normalized for differences in citation practices between scientific fields. For the purpose of this field normalization, about 4000 fields are distinguished. These fields are defined at the level of individual publications. Using a computer algorithm, each publication in Web of Science is assigned to a field based on its citation relations with other publications. By default, the Leiden Ranking ranks universities based on either the size-dependent P(top 10%) indicator or the size-independent PP(top 10%) indicator.

It should be noted that the TCS, MCS, TNCS, and MNCS indicators are not available on the main ranking page. To view these indicator, click on the name of a university, after which a page will be opened that presents an overview of all bibliometric statistics for the university. This overview also includes the TCS, MCS, TNCS, and MNCS indicators.

## Collaboration indicators

The following indicators of scientific collaboration are provided in the Leiden Ranking:

- *P(collab)* and *PP(collab)*. The number and the proportion of a university's publications that have been co-authored with one or more other organizations.
- *P(int collab)* and *PP(int collab)*. The number and the proportion of a university's publications that have been co-authored by two or more countries.
- *P(industry)* and *PP(industry)*. The number and the proportion of a university's publications that have been co-authored with one or more industrial partners.
- *P(<100 km)* and *pp(<100 km)*. The number and the proportion of a university's publications with a geographical collaboration distance of less than 100 km, where the geographical collaboration distance of a publication equals the largest geographical distance between two addresses mentioned in the publication's address list.
- *P(>5000 km)* and *PP(>5000 km)*. The number and the proportion of a university's publications with a geographical collaboration distance of more than 5000 km.

## Core publications

The Leiden Ranking does not simply take into account all publications in Web of Science. Instead, the ranking is based on so-called core publications, which are a subset of all publications in Web of Science. Core publications are publications in international scientific journals in fields that are suitable for citation analysis.

In order to be classified as a core publication, a publication must satisfy the following criteria:

- The publication has been written in English.
- The publication has one or more authors. (Anonymous publications are not allowed.)
- The publication has not been retracted.
- The publication has appeared in a core journal.

The last criterion is a very important one. In the Leiden Ranking, a journal is considered a core journal if it meets the following conditions:

- The journal has an international scope, as reflected by the countries in which researchers publishing in the journal and citing to the journal are located.

- The journal has a sufficiently large number of references to other core journals, indicating that the journal is in a field that is suitable for citation analysis. Many journals in the arts and humanities do not meet this condition. The same applies to trade journals and popular magazines.

In the calculation of the Leiden Ranking indicators, only core publications are included. Excluding non-core publications ensures that the Leiden Ranking is based on a relatively homogeneous set of publications, namely publications in international scientific journals in fields that are suitable for citation analysis. Field-normalized impact indicators such as P(top 10%) and PP(top 10%) become more accurate by excluding non-core publications. It should be emphasized that non-core publications are excluded not because they are considered less important than core publications. Non-core publications may have an important scientific value. About one-sixth of the publications in Web of Science are excluded because they have been classified as non-core publications.

## Counting method

The impact indicators in the Leiden Ranking can be calculated using either a full counting or a fractional counting method. The full counting method gives equal weight to all publications of a university. The fractional counting method gives less weight to collaborative publications than to non-collaborative ones. For instance, if the address list of a publication includes five addresses and two of these addresses belong to a particular university, the publication has a weight of 2 / 5 = 0.4 in the calculation of the impact indicators for this university. The fractional counting method leads to a more proper field normalization of impact indicators and therefore to fairer comparisons between universities active in different fields. For this reason, fractional counting is the preferred counting method for the impact indicators in the Leiden Ranking. Collaboration indicators are always calculated using the full counting method.

## Trend analysis

To facilitate trend analyses, the Leiden Ranking provides statistics not only based on publications from the period 2010–2013, but also based on publications from four earlier periods: 2006–2009, 2007–2010, 2008–2011, and 2009–2012. The statistics for the different periods are calculated in a fully consistent way. For each period, citations are counted until the end of the first year after the period has ended. For

instance, in the case of the period 2006–2009 citations are counted until the end of 2010, while in the case of the period 2010–2013 citations are counted until the end of 2014.

## Stability intervals

Stability intervals aim to provide some insight into the uncertainty in bibliometric statistics. A stability interval indicates a range of values of an indicator that are likely to be observed when the underlying set of publications changes. For instance, the PP(top 10%) indicator may be equal to 15.3% for a particular university, with a stability interval ranging from 14.1% to 16.5%. This means that the PP(top 10%) indicator equals 15.3% for this university, but that changes in the set of publications of the university may relatively easily lead to PP(top 10%) values in the range from 14.1% to 16.5%. The Leiden Ranking employs 95% stability intervals constructed using a statistical technique known as bootstrapping.

# More information

More information on the Leiden Ranking methodology can be found in a number of publications by CWTS researchers. An detailed discussion of the Leiden Ranking is presented by Waltman et al. (2012). This publication relates to the 2011/2012 edition of the Leiden Ranking. Although not entirely up-to-date anymore, the publication still provides a lot of relevant information on the Leiden Ranking. The algorithmic approach taken in the Leiden Ranking to define scientific fields is described in detail by Waltman and Van Eck (2012). Field normalization of impact indicators based on algorithmically defined fields is studied by Ruiz-Castillo and Waltman (2014). The methodology adopted in the Leiden Ranking for identifying core publications and core journals is outlined by Waltman and Van Eck (2013a, 2013b). Finally, the importance of using fractional rather than full counting in the calculation of field-normalized impact indicators is explained by Waltman and Van Eck (2015).

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., Van Eck, N.J., Van Leeuwen, T.N., Van Raan, A.F.J., Visser, M.S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419-2432. DOI: 10.1002/asi.22708.

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378-2392. DOI: 10.1002/asi.22748.

Waltman, L., & Van Eck, N.J. (2013a). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, *96*(3), 699-716. DOI: 10.1007/s11192-012-0913-4.

Waltman, L., & Van Eck, N.J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*(4), 833–849. DOI: 10.1016/j.joi.2013.08.002.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Journal of Informetrics, 9(1), 102–117. DOI: 10.1016/j.joi.2014.11.010.

Waltman, L., & Van Eck, N.J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. Journal of Informetrics, 9(4), 872–894. DOI: 10.1016/j.joi.2015.08.001.