U-Multirank 2017 bibliometrics: information sources, computations and performance indicators

Center for Science and Technology Studies (CWTS), Leiden University

(CWTS version 16 March 2017)

==========

1 Information sources: research publications and patents

1.1 *Web of Science* database

All bibliometric scores are based on information extracted from publications that are indexed in the *Web of Science - Core Collection* database (*Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts & Humanities Citation Index*). CWTS operates this WoS database under a commercial license agreement with Thomson Reuters.

The WoS contains some 14,000 active sources, both peer-reviewed scholarly journals and conference proceedings. The underlying bibliographic information relates to publications classified as 'research article' and 'review article'. The WoS database is incomplete (there are many thousands more science journals worldwide) and it is biased in favor of English-language. Hence, there will always be missing publications. WoS-based bibliometric data are never comprehensive and fully accurate; scores are therefore always estimates with a margin of statistical error.

Nonetheless the WoS is currently one of the two best sources, covering worldwide science across all disciplines. The only possible alternative database, Elseviers' *Scopus* database, has more or less the same features. All in all, one may expect comparable bibliometric results from both databases, especially at higher aggregate levels.

The WoS-indexed publications in Arts and Humanities (A&H) journals have not been included in the three citation-based indicators: (i) mean normalized citation score, (ii) top 10% most frequently cited publications, and (iii) interdisciplinarity indicator. There are three reasons: (1) the citation frequency counts are often zero or low; (2) citation patterns and counts tend to be much more affected by journal- or sub-field specific characteristics; (3) the relatively low level of validity of WoS-indexed peer-reviewed A&H journals as fully representative publication outlets of all research activities in these research disciplines.

The compounded effect of these three constraints is the high likelihood of unreliable and biased outcomes. In combination, the numbers of citations are usually too low to ensure representative, reliable and statistically robust citation-based indicators. Especially in those cases where a *higher*

education institute (HEI) produces low numbers of A&H publications, some of which happen to be (highly) cited, this will give an overly positive view of the HEI's true citation impact in such fields.

Note that publications in arts and humanities are included in all publication-output based indicators – if only to reflect the fact that an HEI is actively engaged in these domains.
For further general information about the Web of Science: http://thomsonreuters.com/products_services/science/science_products/scholarly_research_analysis/research_discovery/web_of_science

1.2 Subject fields

The field-based rankings within U-Multirank are related to fields of science, each of which are defined by collections of peer-reviewed scholarly journals. These journal collections are derived from Thomson Reuters/Clarivate Analytics classification system of *Subject Categories* (SCs). Each WoS-indexed journal is assigned to one or more SCs, according to the general (multi-)disciplinary contents of its publications. There are about 250 SCs in the current system. The four fields that are new within U-Multirank's 2017 edition were defined as follows in the CWTS WoS database:

U-Multirank subject field	**Thomson Reuters/ Clarivate Analytics subject category**
Chemical engineering	Engineering, Chemical
Civil engineering	Engineering, Civil
Industrial engineering	Engineering, Industrial
Economics	Economics and Agricultural Economics & Policy

Please consult the following websites for more information about the SCs:
http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D;
http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=SS.

1.3 *PATSTAT* database

PATSTAT is a database produced by the European Patent Office (EPO) that contains bibliographical data relating to more than 90 million patent documents from more than 100 leading industrialised and developing countries. Patent publications usually contain references to other patents and sometimes also to other 'non-patent' literature sources. A major part of these *non-patent references* (NPRs) are citations to scholarly publications published in WoS-indexed sources. The NPRs are the so-called 'front page citations'. These citations are mainly provided by the patent applicant(s) or by the patent examiner(s) during the search and examination phases of the patent application process.

The citing patents were clustered by using the 'simple patent family' concept – that is groupings of patent publications containing all equivalent, in legal sense, patent documents. A simple patent family therefore addresses one single 'invention'. Each patent family contains at least one EP patent (published by EPO, the *European Patent Office*) or a WO patent (published by WIPO -*World Intellectual Property Organization*), AND at least one patent published by USPTO, the *US Patent and*

Trademark Office. All NPRs within each family were de-duplicated. Each NPR is therefore counted only once per family. The NPRs were matched against the bibliographical records in the WoS. Our current information indicates that the majority of the WoS records are identified.

The patent database used to collect the NPRs from is the Spring 2016 version of the EPO Worldwide Patent Statistical Database (PATSTAT). CWTS operates on an EPO-licensed version of PATSTAT.

2 Technical specifications: data collection, computations, definitions and delineations

2.1 Data preprocessing

The bibliometric indicators are applied to two groups of institutions: (a) the largest 750 universities that are included in the 2016 edition of the CWTS *Leiden Ranking* (6); (b) the main institutions that registered for the U-Multirank. These latter institutions are identified and delineated by CWTS through processing all available author affiliate address information in the publication. CWTS cleans, harmonizes and augments this source of information. The processing involved a mix of sophisticated pattern recognition software, manual checks and corrections, and extensive usage of a CWTS thesaurus of institutional name variants which includes misspellings, acronyms and truncations. For practical and budgetary reasons, this data cleaning and disambiguation work was done 'top down' by CWTS without consulting the HEIs subjected to this process. Some institutions provided input to CWTS on frequently occurring name variants of their originations, which was duly and fully processed.

A key challenge in the delineation and definition of each main institution is the handling of publications originating from closely affiliated research institutes and associated hospitals. 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 and affect 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.

U-Multirank follows the allocation procedure applied by CWTS for its Leiden Ranking, which distinguishes three different types of university-affiliated institutions: component; joint research facility or organization; associated organization.[1] In the case of components the affiliated institution is actually part of the 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 are examples of components. All teaching and research tasks in the field of medicine that were traditionally the

[1] This paragraph is largely copied from the explanatory text on the *Leiden Ranking* website (accessed on 18-2-2015).

responsibility of the universities have been delegated to these separate organizations that combine the medical faculties and the university hospitals. 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 both examples of this type of affiliated institution. As for associated organizations, the third category, take the case of *Addenbrooke's Hospital* in Cambridge (UK), an organization associated with *Cambridge University.* Publications mentioning only the Addenbrooke's Hospital are not counted as publications from *Cambridge University. Only publications explicitly mentioning the Cambridge University or one of its components are included.* However, as many *Addenbrooke's Hospital* affiliations appear within publications alongside another address referring to Cambridge University, these publications will in fact be attributed to to Cambridge University.

Organisational sub-units that registered for participation in U-Multirank - such as individual faculties, schools, departments, or institutes – were excluded from the CWTS bibliometric data collection. Mainly because many of these sub-units are extremely difficult to delineate from the parent organization because the verification of author address information, and additional data collection, requires extensive input and feedback from knowledgeable representatives of the sub-units.

These data processing complexities also occurred in the Tehran branches of *Islamic Azad University* which was registered in U-Multirank. The available address information proved of insufficient quality to consolidate these branches as separate institutions.. Hence, no reliable bibliometric data could be provided for these entities. However, for *Islamic Azad University, Najafabad Branch* which is the only branch located in *Najafabad*, information could be provided as the address information was sufficiently reliable.

2.2 Indicators, metrics and computational issues

General background

All the CWTS-generated bibliometric indicators presented in this section are either fully or partially derived from pre-existing generally available indicators, or based on prior CWTS ideas or research that occurred outside the U-Multirank project. In some cases the bibliometric scores on these indicators were derived from prior CWTS-developed data-processing routines or computational algorithms, or modified/upgraded versions thereof.

The WoS-based bibliometric scores relate to the publication years 2012 up to and including 2016, as a single measurement window, with the exception of the 'Patent citations to research publications' metrics.

Leiden Ranking

The bibliometric indicators in U-Multirank are closely related to those in the Leiden Ranking. The main difference between both is the fact that the Leiden Ranking is based on WoS-indexed 'core research publications' in international peer-reviewed scientific journals. Publications in other WoS-indexed sources (national scientific journals, trade journals, and popular magazines) are not included. The same applies to research publications in languages other than English. Also publications in journals that are not well-connected, in terms of citation links, to other journals are left out. (These are mainly, but not exclusively, journals in arts and humanities fields of science). For a brief explanation of the idea of core publications, see *http://www.leidenranking.com/methodology/indicators#core-journals*. In contrast, U-Multirank includes all WoS-indexed sources and publications (although for some indicators A&H publications are left out – see below).

Full counting or fractional counting

The bibliometric indicators fall into two groups, depending on the counting scheme (fractional or full) and the coverage of the arts and humanities research publications (included or excluded). The following indicators use full counting and include the arts and humanities: Research publication output; International co-publications; Regional co-publications; Co-publications with industrial partners. Three of other indicators exclude arts and humanities publications and use a fractional counting scheme: Interdisciplinary research score; Mean normalized citation score; Frequently cited publications. Finally, Patent citations to research publications also excludes arts and humanities publications, but it uses full counting. We refer to the recent publication by Waltman and Van Eck (2015; http://arxiv.org/abs/1501.04431) to justify our use of fractional counting for some of these indicators.

Publication output threshold values

Measurement processes and bibliometric data that are based on low numbers of publications are more likely to suffer from 'small size effects', where small (random) variations in the data might lead to very significant deviations and discrepancies. Higher education institutions (HEIs) with a only few publications in WoS-indexed sources should therefore not be described by WoS-based indicators (alone). To prevent this from happening threshold values were implemented. No bibliometric scores will be computed for the institutional ranking if the institution produced less than 50 WoS-indexed publications during the years 2012-2015. This count is based on a full counting scheme where each publication is allocated in full to every main institution mentioned in the publication's author affiliate addresses. Where the institutional ranking relates to all research publication output, irrespective of the field of science, the three field-based rankings relate to specific fields. The list of fields and their delineation in the WoS database is explained in section 1.2. The lower publication output threshold for each field is set at 20 (full counted) WoS-indexed publications. Both these lower cut-off points were approved in December 2013 by U-Multirank's Advisory Board.

Note that these same two lower thresholds also apply to the indicator 'Patent citations to research publications', which applies an extended 10-year time-period for computational reasons (see below)

and enables the accumulation of larger numbers of publications. Hence, in some cases data is provided for this particular indicator but not for the others.

3 Bibliometric performance indicators

3.1 Research publication output

Indicator of: volume of research activity
Metric: frequency count of research publications

Background and specification: the number of WoS-indexed publications produced by a main institution reflects the magnitude of international-level research activity. The publication frequency count data are based on a whole counting system, where each publication is assigned in full to every main organization mentioned in its author affiliation list. Publication output counts do not necessarily reflect the volume of in-house research capacity, due to the existence of significant disciplinary differences between publication propensities and the output-boosting effect of research cooperation with external institutional partners.

3.2 Interdisciplinary research score

Indicator of: knowledge usage from different scientific disciplines
Metric: share of publications within the field's top 10% publications with the highest interdisciplinarity scores

Background and specification: the frontiers of science are often at the edge of disciplines – those dynamic domains of cross-fertilization where insights, ideas and information from other disciplines lead to new understanding and scientific breakthroughs. The term 'interdisciplinarity' is used to capture this feature of a HEI's research profile. Our measure of the average interdisciplinarity of the publications of an institution aims to capture the diversity in the knowledge sources of publications. The interdisciplinarity score of a single publication is determined based on the references ('citations') within that publication to other WoS-indexed publications. The more a publication refers to publications belonging to different fields of science, and the larger the cognitive distance between these fields, the higher the interdisciplinary score of that publication will be. More precisely, the interdisciplinarity score of a publication equals the average, calculated over all pairs of cited publications, of the distance between the fields to which the cited publications belong.

The distance between two fields is determined based on citation relationships between fields. The more two fields cite to the same fields (as calculated using the so-called cosine measure), the smaller the cognitive distance between the two fields. After the interdisciplinarity scores of all publications in *Web of Science* database in the period of analysis have been calculated, we refer to the top 10%

research publications with the highest interdisciplinarity score are as 'highly interdisciplinary' publications. The results of our sensitivity analyses indicates that the ranking is relatively insensitive *to* the choice of the percentage that are classified as interdisciplinary; choice of the percentage that is classified as interdisciplinary (either 10% or another percentile). In order to obtain the interdisciplinarity score of an institution, its proportion of interdisciplinary publications is calculated across all fields of science collectively. Mathematically, the interdisciplinarity score of an individual publication can be written as:

$$I^{\text{pub}} = \frac{1}{m^2}\sum_{i,j} d_{ij}$$

where *m* denotes the number of references in the publication to other WoS-indexed publications and where d_{ij} denotes the distance between the field of reference *i* and the field of reference *j*. The distance d_{ij} equals 0 if reference *i* and reference *j* are in the same field. The maximum possible value of d_{ij} is 1. The interdisciplinarity score of an institution equals the proportion of the publications of the institution that are regarded as highly interdisciplinary, or in other words, the proportion of the publications of the institution that belong to the top 10% publications with the highest interdisciplinarity score in their field per year. In mathematical terms, this can be written as

$$I^{\text{inst}} = \frac{1}{n}\sum_{k} \#\left(I_k^{\text{pub}} \geq I_{\text{threshold}}^{\text{pub}}\right)$$

where *n* denotes the number of publications of the institution, I_k^{pub} denotes the interdisciplinarity score of publication *k*, and $I_{\text{threshold}}^{\text{pub}}$ denotes the minimal interdisciplinarity score a publication must have in order to belong to the top 10% publications with the highest interdisciplinarity score. We refer to Porter and Rafols (2009)[2] for a further discussion of the above approach for measuring interdisciplinarity, in particular the approach for calculating the interdisciplinarity score of an individual publication. This publication also explains in detail how the distance between two fields can be calculated using the cosine formula.

3.3 Percentage of international co-publications

Indicator of: research cooperation with partners in other countries
Metric: share of research publications with at least one author affiliate address located in another country

Background and specification: the percentage of the publications, within a HEI's research publication output, with one or more co-authors publishing with an affiliate address in another country. Each

[2] Porter, A.L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, 81(3), 719–745.

international co-publication is assigned in full to all main organizations listed in those addresses. These co-publication counts are slightly affected by (temporarily employed) researchers with one or more appointments abroad. The international co-publication propensity is discipline-specific; it is relatively high in the natural sciences; relatively low in the social sciences, and extremely low in the arts and humanities fields.

3.4 Percentage of regional co-publications

Indicator of: research cooperation with partners in the local region
Metric: share of publications with two or more author addresses located within a 50 kilometer radius of the university

Background and specification: captures the extent to which HEIs collaborate and co-publish with external institutional research partners located at close proximity. The metric is defined in terms of physical distance (measured in kilometers) between the HEI and its partner. The local 'region' is defined as a 50 km radius around the city center of the university's main location. A publication is considered to represent a short distance collaboration for a particular institution if, apart from the address of this institution, at least one other address is mentioned in the address list of the publication and if this other address is within 50 km of the address of the institution of interest.

3.5 Percentage of co-publications with industrial partners

Indicator of: research cooperation with business enterprises
Metric: share of publications with at least one author affiliate address referring to a for-profit business company

Background and specification: percentage of an institution's research publications with co-authors employed by 'industry' - delineated as for-profit business companies, but excluding private-sector education institutions and hospitals/clinics. Most of the enterprises therefore operate in manufacturing industries. The share of co-publications with industry is discipline-specific and depends on the type of HEI; it is relatively high in industry-relevant fields within the engineering sciences and life sciences, and among universities of technology. It is important to note that this indicator may also comprise of cases where staff have (temporary or permanent) dual appointments of affiliations both within a university and a business companies, or where former PhD students, recently employed by industry, still publish about their academic work under their previous university address.

3.6 Mean normalized citation score

Indicator of: international scientific impact
Metric: average citation impact of research publications corrected for field-specific characteristics worldwide

Background and specification: absolute number of citations received by a publication is often highly dependent on the field of science, the topic of the publication, and sometimes even the source in which it was published. Proper citation counting needs to take this into account, in order to compare across research domains and different types of HEIs. The average number of citations, from other WoS-indexed publications, to publications of an institution, normalized at the global level for the field and the year in which a publication appeared. This normalization aims to correct for differences in citation characteristics between publications from different fields and different years.

Citations are counted up to and including the fourth quarter of 2016, where author self-citations are ignored in the computations. The fields we use for normalization are identical to the *Leiden Ranking* methodology (www.leidenranking.com) which are based on clusters of interlinked research publications. We adopt a fractionated counts in the citation analysis, where a cited publication is allocated to an institution in proportion to the number of times the main organization is mentioned in author affiliate addresses.

3.7 Percentage of highly cited publications

Indicator of: high-level scientific impact
Metric: share of research publications within the top 10% most highly cited of their field worldwide

Background and specification: citation distributions are highly skewed – the top 10% most highly publication collect on average some 50-60% of all citations worldwide. This indicator captures the share of a HEI's publication output that belongs to the top 10% most frequently cited per field worldwide. This measure is occasionally introduced as an indicator of 'international research excellence': HEIs with well over 10% of their publications in this top percentile are among the top research institutes worldwide. Note that these very highly cited publications are very often internationally co-authored publications. Citations are counted up to and including the third quarter of 2016, where author self-citations are ignored in the computation. Similarly to the Mean normalized citation score (see above), the citation counts are based on a fractional counting scheme.

3.8 Patents awarded

Indicator of: knowledge transfer and technological development
Metric: number of patents awarded to inventors working in the university

Background and specification: knowledge generated at universities may flow through different channels to actors outside the university. The number of patents is an established measure of technology transfer, as it indicates the degree to which inventions made in academic institutions may be transferred to economic actors for further industrial / commercial development. The number of patents is one of the indicators reflecting the university *potential* transfer of knowledge with

commercial application. However, the existence of patents it is not a synonym of actual knowledge transfer. Additional information, for instance on patent licenses, is required to determine whether the knowledge disclosed in a patent document has actually flowed and has been used by actors outside academia.

The indicator on the number of patents awarded has been calculated for each university considering the number of patents granted at the US patent office (USPTO) and/or the European Patent Office (EPO) for patents applied between 2004 and 2013. It is frequent that the same invention is protected in more than one patent office (e.g. USPTO and EPO), therefore in order to avoid double counting we rely on the INPADOC patent families so that the number of patents equals to the number of INPADOC patent families with at least one granted patent at the USPTO and/or the EPO. Many universities apply for their patents under the name of specific units created to manage the transfer of technology at the universities (e.g. *ISIS Innovation* at University of Oxford). In the collection of patents, all the known technology transfer units have been considered. However, it is possible that we are not aware of all the possible names under which the patents have been applied for, and in some cases the indicator may underrepresent the actual number of patents of the university.

3.9 Industry co-patents

Indicator of: knowledge transfer and technological development
Metric: number of co-patents with industry awarded to inventors working in the university

Background and specification: inventions protected by patents might be developed by universities alone or in collaboration with other partners, for instance companies. These patents reflect a close interaction between the university and the firm which, in principle, takes place at an early stage of the development of the invention protected through the patent. This indicator refers to the number of patents awarded that were applied at the same time by universities and companies, being both mentioned as patent applicants. The underlying methodology to identify these patents is the same used to calculate the indicator on 'Patents awarded'. For the identification of companies, we rely on the existing classification of applicants in PATSTAT.

3.10 Share of research publications cited by patents

Indicator of: impact of scientific research on technological development
Metric: percentage of research publications cited in patented technologies

Background and specification: the percentage of a HEI's research publications that were mentioned in the reference list of at least one international patent – the so-called 'front page' references which are assembled separately from the patent's main text. An 'international patent' is defined here as a patent belonging to a DOCDB patent family, which each of consisting of equivalent patent

publications, describing the same invention, published either by the US Patent and Trademark Office (USPTO), the European Patent Office (EPO) or the World Intellectual Property Organization (WIPO). This indicator reflects the *technological relevance* of scientific research at the HEI, in the sense that it explicitly contributed, in some way, to the development of patented technologies. It does not necessarily reflect the *innovation performance* of HEIs; a patented technological only becomes an innovation when the related product or process is introduced into the marketplace.

Note that not all references in patents reflect a direct link between the 'cited' science and 'citing' technology, and that the cited publications can be co-authored with other organizations. Nonetheless, a relatively large share of cited research publications will reflect that HEIs in the recent past have been, and most likely still are, engaged in research of (future) technological importance. To compensate for the relatively low number of NPRs to WoS-indexed publications, a 10-year time-period was adopted for the citing patents (2005-2014) in order to capture sufficiently large numbers for statistical meaningful comparisons. Because of patent publication delays, PATSTAT's the coverage of the PATSTAT version used (Spring 2016) is incomplete for 2014.

The citation window for the cited research publications is 2005-2014. The number of patents citations to research publications is relatively low; scores on this indicator therefore have to be treated with due care because of statistical error.