Bibliometric Analysis in U-Multirank

Information source: bibliographic records of research publications and patents

Web of Science database

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

The WoS contains some 12,000 sources, mostly peer-reviewed scholarly journals. 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 this source:

http://thomsonreuters.com/thomson-reuters-web-of-science/

Updated: July 21, 2014

Subject Categories

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' 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 within U-Multirank's 2014 edition are defined as follows:

U-Multirank field	Thomson Reuters Subject Category
Business and Management	Business
	Business - Finance
	Management and Planning
Electrical Engineering	Engineering - Electrical & Electronic
	Telecommunications
Mechanical Engineering	Engineering - Mechanical
Physics	Physics - Applied
	Physics - Atomic - Molecular & Chemical
	Physics - Condensed Matter
	Physics - Fluids & Plasmas
	Physics - Mathematical
	Physics - Multidisciplinary
	Physics - Nuclear
	Physics - Particles & Fields

Please consult the Thomson Reuters websites for the list of journals per SC:
http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D

http://science.thomsonreuters.com/cgi-bin/jrnlst/jloptions.cgi?PC=K

PATSTAT database

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 provided by the patent applicant(s) or by the patent examiner(s) during the search and examination phases of the patent application process.

The patent database used to collect the NPRs from is the April 2013 version of the EPO Worldwide Patent Statistical Database (PATSTAT). CWTS operates an EPO-licensed version of PATSTAT.

For further general information about this source:
http://www.epo.org/searching/subscription/raw/product-14-24.html

The period of analysis covers the publication years 2001–2010.

Bibliometric indicators: data collection, computations, and other technical specifications

Data preprocessing

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. HEIs are excluded from a score in the institutional ranking if they produced less than 50 WoS-indexed publications in 2008-2011. Where the institutional ranking relates to all research publication output, irrespective of the field of science, the four field-based rankings related to specifics fields. In most cases one may expect less publications than the total output across all fields. The lower threshold for individual fields was set at 20 WoS-indexed publications. Both cut-off points were approved by the U-Multirank Advisory Board during the December 2013 meeting.

The bibliometric indicators are applied to two groups of institutions: (a) the 500 universities that are included in the 2013/2014 edition of the CWTS Leiden Ranking (www.leidenranking.com/ranking/2013); (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.

Hospitals, clinics and other medical centers are included in the affiliated universities and medical schools. This is standard CWTS procedure in bibliometric analyses at the main organizational level.

Organisational sub-units that registered for participation in U-Multirank - such as individual faculties, schools, departments, or institutes – were excluded from the CWTS bibliometric method. 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.

The WoS-based bibliometric scores relate to the publication years 2008 up to and including 2011, as a single measurement window.

The PATSTAT database was used to extract research publications that are listed in the reference lists of international patents – the *non-patent references* (NPRs). 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 crude estimates indicate that the majority of the WoS records are identified, but accuracy of the matching procedure is still unknown, and is subject of validation studies at CWTS.

To compensate for the relatively low number of NPRs to WoS-indexed publications, a 10-year time-period was adopted (from 2001 to 2010), in order to capture sufficiently large numbers for statistical meaningful comparisons. This time-span aligns with data collection for the patent output indicator, which was done by INCENTIM (Catholic University of Leuven), a U-Multirank partner institute.

The linkage between the NUTS classification system and the WoS database was done my matching the information in the author affiliate addresses in WoS-indexed research publications to the two defining information items of NUTS regions: the city names and/or the associated postal codes. A pre-existing dedicated CWTS-matching algorithm was modified to conduct this task. More than 95% of the research publications with sufficient address information have been unambiguously matched to the corresponding NUTS region in Europe.

Bibliometric indicators

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 development of two indicators have benefitted significantly from U-Multirank related contributions: 'Percentage of regional co-publications' and 'Interdisciplinary research score'.

<u>Research publication output</u>
Indicator: frequency count of publications

Description: 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.

<u>Percentage of international co-publications</u>
Indicator: share of publications with at least one author address located in another country

Description: 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 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.

Percentage of regional co-publications

Indicator: share of publications with two or more author addresses located within the HEI's immediate geographical zone ('region')

Description: This indicator captures the extent to which HEIs collaborate and co-publish with external institutional research partners located at relatively close proximity. The indicator is defined in terms of physical distance between the location of the HEI and its closest located partner – as observed within the author affiliate address list on the same research publication. If the distance is less than 50 kilometers, the partner is considered to be within the same region. HEIs that are located close to a border with another country may have regional research partners that are also international partners. The share of regional partners will tend to be relatively large in major cities and urban agglomerations with a high density of HEIs or public research institutes.

Percentage of co-publications with industrial partners

Indicator: share of publications with at least one author address referring to a for-profit business company

Description: The 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. The share of HEI-industry is positively affected by staff with (temporary) appointments in both sectors of former HEI students, now employed by industry, still also publishing under their former institutional address.

Mean normalized citation score

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

Description: The 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 until the end of 2012, where author self-citations are ignored in the computations.

Percentage of frequently cited publications

Indicator: share of research publications within the most highly cited of their field worldwide

Description: Citation distributions are highly skewed – the top 10% most highly cited publications 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 until the end of 2012, where author self-citations are ignored in the computation.

Interdisciplinary research score

Indicator: extent to which reference lists of publications reflect inputs from different scientific disciplines

Description: 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 activity and output.

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 distance between these fields, the higher the interdisciplinary score of the 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 distance between the two fields. In order to obtain the interdisciplinarity score of an institution, the interdisciplinarity scores of the publications of the institution are averaged.

For background information on technicalities, please consult the research article: Porter, A.L., Ismael Rafols, I., "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time", Scientometrics, 81 (3), 719-745, 2009.

Patent citations to research publications

Indicator: share of research publications cited in selected patent families

Description: The percentage of a HEI's research publications that were mentioned in the reference list of at least one international patent[1] – the so-called 'front page' references which are assembled separately from the patent's main text. This indicator therefore 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.

Note that not all references 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 publications will reflect that HEIs have been, and most likely still are, engaged in research of technological importance. It does not necessarily reflect the *innovation performance* of HEIs; a

[1] 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).

patented technological only becomes an innovation when the related product or process is introduced into the marketplace.

The number of these citations is relatively low; scores on this indicator have to be treated with due care because of statistical error.

performance of HEIs; a patented technological only becomes an innovation when the related product or process is introduced into the marketplace.

The number of these citations are relatively low; scores on this indicator therefore have to be treated with due care because of statistical error.