RISIS RESEARCH INFRASTRUCTURE FOR RESEARCH AND INNOVATION POLICY STUDIES

Report on the content and technical structure of the

CWTS Leiden Ranking 2016

Infrastructure

Leiden University

RISIS "Research infrastructure for research and innovation policy studies"

FP7, Grant agreement no: 313082

Task 6, Workpackage 6, coordinated by AIT Austrian Institute of Technology Gmbh

SEVENTH FRAMEWORK PROGRAMME

Report on the content and technical structure of CWTS Leiden Ranking 2016 (Task 6 of WP6)

Ed Noyons & Clara Calero

*CWTS, Leiden University

Table of contents

1 Basic characteristics 3
2 Information on substantive content of *CWTS Leiden Ranking 2016* 3
2.1 Definition and description of observations 3
2.2 Data acquisition and processing 4
2.3 Information on all variables/indicators 6
2.4 Sectorial, temporal and geographical coverage 6
2.5 Quality and accuracy of data 7
3 Legal issues encountered and access conditions 7
4 Technical structure of *CWTS Leiden Ranking 2016* 8
4.1 Information on the data base system 8
4.2 Technical variable definition 8
4.3 Description of the Entity Relationship Model of *CWTS Leiden Ranking 2016* 9
4.4 Interfaces for access and to other infrastructures 9
5 Further planning of the opening of *CWTS Leiden Ranking* 10

1 Basic characteristics

Name and short description of the infrastructure

CWTS Leiden Ranking is a web-service of a university ranking focusing on output and impact of research. The underlying data are collected and processed from the CWTS version of Web of Science (WOS).

Aim of the database (context of data acquisition)

The CWTS Leiden Raking aims at providing a platform to contribute to the discussion on university rankings. We created this platform to share the standardized way of measuring research output and impact. The measurement regards the performance of universities in a recent period of time (5 years) and is as such cross sectional. The current CWTS Leiden Ranking 2016 data cover the publication years 2010-2015 and citations until 2013.

Legal name of operating organization

Universiteit leiden – Centrum voor Wetenschaps- en technologie Studies (CWTS)

Database location and type of access (access on site, online, etc.)

CWTS Leiden Ranking is online available at http://www.leidenranking.com and onsite at

Leiden University, CWTS

Willem Einthoven building

Wassenaarseweg 62A

2333 AL Leiden

The Netherlands

2 Information on substantive content of *CWTS Leiden Ranking 2016*

2.1 Definition and description of observations

Units and definition of observations

CWTS Leiden Ranking covers information on:

- Universities worldwide

Number of observations

CWTS Leiden Ranking comprises information on ***750 universities.*** These 750 are the most productive ones in 2010-2014 according to data from WoS.

2.2 Data acquisition and processing

Underlying data

CWTS Leiden Ranking is created using bibliographic and citation data from the CWTS version of WoS. The CWTS version is an exact copy of the WoS but with specific enhancements to the data (cleaned addresses and sources, added indicators etc.). The address data (affiliation of authors) of millions of publications is processed and cleaned to identify the most productive universities worldwide.

Data processing and cleaning

The CWTS WoS version enables a continuous processing and cleaning of new publication (hence address) data. The process for cleaning universities names regards an hierarchical approach in which first country names are cleaned, then city names and regions before the cleaning of university names starts. Within each city and region variants are identified and harmonized.

Identification of universities

The criteria that have been adopted to define universities for the Leiden Ranking are not very formal. 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. The focus of the Leiden Ranking on scientific research certifies that the institutions included in the Leiden Ranking 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. These missions and responsibilities in turn are strongly linked to the national and regional academic systems in which universities operate. 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. Again, 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. University of California) were split up into separate universities. The higher education sector in France, like in many other countries, has gone through many reorganizations in recent years. Many French institutions of higher education have been grouped together in Pôles de Recherche et d'Enseignement Supérieur (PRES), 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 2013 have been taken into account. For example, in the Leiden Ranking 2016, the University of Lisbon which merged with the Technical University of Lisbon in 2013 encompasses all publications assigned to the old University of Lisbon as well as the publications previously assigned to the Technical University of Lisbon.

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 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.

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 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 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.

The third type of affiliated institution is the associated organization which is more loosely connected to the 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 in particular is of substantial consequence as 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 2016, 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 type of affiliations are credited to the contribution of that particular university to the publication in the fractional counting method.

2.3 Information on all variables/indicators

CWTS Leiden Ranking indicators for universities

- P: number of publications of a university
- TCS (total citation score): the total amount of citations received (self-citations excluded)
- MCS (mean citation score). The average number of citations of the publications of a university.
- TNCS (total normalized citation score): Total number of citations received normalized by field and publication year.
- MNCS (mean normalized citation score). The average number of citations of the publications of a university, normalized for field differences and publication year. An MNCS value of two for instance means that the publications of a university have been cited twice above world average.
- PP(top 10%) (proportion of top 10% publications). The proportion of the publications of a university that, compared with other publications in the same field and in the same year, belong to the top 10% most frequently cited
- PP(collab) (proportion of interinstitutional collaborative publications). The proportion of the publications of a university that have been co-authored with one or more other organizations.
- PP(int collab) (proportion of international collaborative publications). The proportion of the publications of a university that have been co-authored by two or more countries.
- PP(UI collab) (proportion of collaborative publications with industry). The proportion of the publications of a university that have been co-authored with one or more industrial partners. For more details, see University-Industry Research Connections 2013.
- PP(<100 km) (proportion of short distance collaborative publications). The proportion of the publications of a university 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.
- PP(>1000 km) (proportion of long distance collaborative publications). The proportion of the publications of a university with a geographical collaboration distance of more than 1000 km.

2.4 Sectorial, temporal and geographical coverage

Information on the sectorial classifications used

The publications are distributed over the following main fields:

- All fields
- Cognitive and health sciences
- Earth and environmental sciences

- Life sciences
- Mathematics, computer science, and engineering
- Medical sciences
- Natural sciences
- Social sciences

The above fields have been defined using a unique bottom-up approach. Traditionally, fields are defined as sets of closely 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 particular field. The seven broad fields of science listed above have been defined at the level of individual publications rather than at the journal level. Using a computer algorithm, each publication in the Web of Science database has been assigned to one of these seven fields. This has been done based on a large-scale analysis of hundreds of millions of citation relations between publications. More information about the main fields and how they are selected via http://www.leidenranking.com.

Information on the temporal coverage used

The CWTS Leiden Ranking 2016 data cover the publication years 2010-2014 and citations until 2015.

Information on the geographical coverage and classifications used

The CWTS Leiden Ranking 2016 data cover 750 most productive universities worldwide. In the current version they represent all regions Africa, Asia, Europe, North America, South America, and Oceania. Within these regions, the ranking covers 49 countries.

The classification used to normalize indicators and distinct main fields is an inhouse publication classification. More information about this classification and main fields at http://www.leidenranking.com or http://www.cwts.nl.

2.5 Quality and accuracy of data

Estimation of data quality issues with respect to data acquisition, reliability of retrieving system

It is important to highlight that the assignment of publications to universities is not free of errors. There are generally two types of errors: 'false positives', which are publications that have been assigned to a university when they do not in fact belong to that university, and 'false negatives', which are publications that have not been assigned to a university when they should in fact have been. Considerably more false negatives than false positives should be expected, especially since the 5% least frequently occurring addresses in the database may not have been manually checked. This can be considered a reasonable upper bound for errors, since the majority of these addresses are probably non-university addresses.

3 Legal issues encountered and access conditions

Legal issues concerning access of the database

The data disclosed in the CWTS Leiden Ranking 2016 (as well as previous editions) is publically available at http://www.leidenranking.com. The data can be used interactively through a web interface and is also available by XLS download.

For more detailed analysis an onsite visit is needed to access the underlying data of the CWTS Leiden Ranking databases. For this we host a guest account and an Non-Disclosure Agreement (NDA) needs to be signed.

Owner of CWTS Leiden Ranking raw data

Universiteit Leiden, CWTS is the owner of the CWTS Leiden Ranking data.

Copyright holder for underlying WoS data is Thomson Reuters.

Current practice for opening up of the database to external users

The data of the CWTS Leiden Ranking 2016 (and previous editions) is publically available via http://www.leidenranking.com. A downloadable XLS version is also available of all data.

Legal necessities for potential opening procedures

Access to the underlying CWTS WoS data is possible onsite at CWTS. For this a guest account and signed NDA will be arranged.

4 Technical structure of *CWTS Leiden Ranking 2016*

4.1 Information on the data base system

Current data base system used

CWTS Leiden Ranking is processed and stored in a Microsoft SQL server dedicated web database.

Planned future technical changes concerning data base system

None

4.2 Technical variable definition

Labelling and data type of all variables

The indicators included are as listed in section 3.2. The indicators with the extension '_lb' and '_ub' regard the lower bound and upper bound values used to define the stability intervals. The following list describes the labelling of the variables, data type is given in brackets:

- University (char)
- Country (char)
- Field (char)
- Frac_counting (0 or 1)
- P (num)
- TCS (num)
- TNCS (num)
- P_top (num)
- P_collab (num)
- P_int_collab (num)
- P_UI_collab (num)
- P_short_dist_collab (num)
- P_long_dist_collab (num)
- MCS (num)
- MCS_lb (num)
- MCS_ub (num)

- MNCS (num)
- MNCS_lb (num)
- MNCS_ub (num)
- PP_top (perc)
- PP_top_lb (perc)
- PP_top_ub (perc)
- PP_collab (perc)
- PP_collab_lb (perc)
- PP_collab_ub (perc)
- PP_int_collab (perc)
- PP_int_collab_lb (perc)
- PP_int_collab_ub (perc)
- PP_UI_collab (perc)
- PP_UI_collab_lb (perc)
- PP_UI_collab_ub (perc)
- PP_short_dist_collab (perc)
- PP_short_dist_collab_lb (perc)
- PP_short_dist_collab_ub (perc)
- PP_long_dist_collab (perc)
- PP_long_dist_collab_lb (perc)
- PP_long_dist_collab_ub (perc)

4.3 Description of the Entity Relationship Model of *CWTS Leiden Ranking 2016*

Definition of single tables

CWTS Leiden ranking 2016 consists of one table (see previous section). The underlying data consists of three tables:

List of Universities:

ID	num
Name	char

Universities_publications

ID	num
Pub_id	char

Publication indicators

Pub_id	char
Indicator A	num
Indicator B	num
Etc.	num

Relation between the tables via unique identifiers

The table *List of universities* is connected to *universities_publications* via ID. The *Universities_publications* table is connected to the table *Publication indicators* via pub_id.

4.4 Interfaces for access and to other infrastructures

Technical information on interfaces with other infrastructures (e.g. web interface for data search. etc.)

The CWTS Leiden Ranking 2016 and previous editions is available via http://www.leidenranking.com.

5 Further planning of the opening of *CWTS Leiden Ranking*

Document concrete steps towards opening of the respective dataset

The results of the CWTS Leiden Ranking are already online. Underlying tables are available at CWTS onsite. A guest account at the CWTS and a signed NDA is needed to access these tables.

The underlying address data is being updated continuously. This data is also available at CWTS onsite under the same conditions (guest account and signed NDA).

Necessary updates and/or technical changes

The CWTS Leiden Ranking is updated on a yearly basis. In April an update is launched online (website) containing publications until the year before the preceding and citations until the preceding year.

During the RISIS period, new entries will be included in the CWTS Leiden Ranking. We anticipate to include Public Research Organizations (PROs) in the 2017 edition.

Changing legal conditions for accessing the dataset or parts of the dataset

The legal access conditions are expected to remain the same throughout.