RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

228

A Brief Guide for the Creation of Author-specific Citation Metrics and Publication Data Using the Thomson Reuters Web of Science and Scopus Databases

Frank Mueller-Langer
Michael Gerstenberger
Julian Hackinger
Benjamin Heisig

December 2013

© SCIVERO Press

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

The RatSWD Working Paper Series is edited by:

Chair of the RatSWD (2007/2008 Heike Solga; since 2009 Gert G. Wagner)

Managing Director of the RatSWD (Denis Huschka)

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

A Brief Guide for the Creation of Author-specific Citation Metrics and Publication Data Using the Thomson Reuters Web of Science and Scopus Databases

Frank Mueller-Langer[1], Michael Gerstenberger, Julian Hackinger and Benjamin Heisig

All authors: Max Planck Institute for Innovation and Competition, Munich Center for Innovation and Entrepreneurship Research (MCIER)

Abstract

The objective of this guide is twofold. First, it shall enable interested readers to understand and reproduce the process of collecting author-specific citation metrics and publication data from the Thomson Reuters Web of Science and Scopus databases that is adopted in Andreoli-Versbach and Mueller-Langer (forthcoming). Second, it presents the problems faced during the data collection process and the refined method of data collection we adopt to address related concerns. Thereby, it may serve interested readers as a guideline to accurately and efficiently retrieve citation metrics and publication information from Thomson Reuters Web of Science and Scopus in similar endeavors.

Keywords: Citation metrics, publication data, Thomson Reuters Web of Science, Scopus
JEL classification: C80

1 Introduction

The database provided by Thomson Reuters "Web of Science" (WOS) is frequently used to retrieve author-specific citation metrics, publication data and journal impact factors.[2] Researchers and journals are often ranked on the basis of the output generated from WOS and Scopus.

[1] Contact author: Dr. Frank Mueller-Langer, Munich Center for Innovation and Entrepreneurship Research, Max Planck Institute for Innovation and Competition, Marstallplatz 1, 80539 Munich, Germany. We gratefully acknowledge financial support from the German Research Foundation (DFG) under the European Data Watch Extended Project (EDaWaX). EDaWaX is a joint project of the German National Library of Economics (ZBW), the German Data Forum (RatSWD) and the Munich Center for Innovation and Entrepreneurship Research (MCIER) at the Max Planck Institute for Innovation and Competition supported by the International Max Planck Research School for Competition and Innovation (IMPRS-CI). We thank Hermann Schier and Gert G. Wagner for their valuable comments and suggestions.

[2] See Gaulé and Maystre (2011), Hitt and Greer (2012), Furman and Stern (2011) and McCabe (2002) as well as the literature cited therein.

The objective of this guide is twofold. First, it shall enable interested readers to understand and reproduce the process of collecting author-specific citation metrics and publication data from the WOS[3] and Scopus[4] databases that is adopted in Andreoli-Versbach and Mueller-Langer (forthcoming). Second, it shall present the problems that we face during the data collection process and propose a refined method of data collection that we adopt to address these problems. Thereby, it may serve interested readers as a guideline to efficiently retrieve citation metrics and publication information from WOS in similar endeavours.

The analysis in Andreoli-Versbach and Mueller-Langer (forthcoming) is based on a hand-collected sample of 488 randomly chosen authors in the field of empirical economic research. 388 (100) researchers are affiliated with one of the top-100 economics departments (top-50 business schools) according to the Shanghai Ranking 2011 in Economics and Business[5] (the Financial Times Global MBA Ranking 2011[6]). Our objective is to generate a complete list of articles that the authors under study have published in WOS-listed economics and business journals, e.g., title, journal name, year, volume, issue, field of research etc., as well as author-specific citation metrics, e.g., h-index, total citations, average citations per paper, citations per year etc., and journal-impact factors.[7] To illustrate, this information allows the analysis of the question of whether higher-quality authors as measured by citation metrics, e.g., citations per article, are (ceteris paribus) more likely to voluntarily share their research data with the scientific community.

In a nutshell, we adopt the following approach. First, we download the latest CVs of the authors under study from their personal websites or from the website of the current institution they are affiliated with to get a complete list of publications. Second, we retrieve the relevant author-specific publication data and citation metrics from WOS by searching for the name of the respective author, extract the output (being the set of articles from a single author published in WOS-listed journals) and convert the output to a spreadsheet. Finally, we merge the 488 single spreadsheets into one single file.[8]

In the following, we outline the problems that we face along the way and how we address related concerns in order to ensure the integrity and consistency of the data creation process adopted in Andreoli-Versbach and Mueller-Langer (forthcoming).

[3] The WOS database is available at http://apps.webofknowledge.com/ (last visited October 15, 2013)
[4] The Scopus database is available at http://www.scopus.com/ (last visited October 15, 2013).
[5] The Shanghai Ranking is available at http://www.shanghairanking.com/SubjectEcoBus2011.html (last visited October 15, 2013).
[6] The FTA Global MBA Ranking is available at http://rankings.ft.com/businessschoolrankings/global-mba-rankings-2011 (last visited October 15, 2013).
[7] For a critical discussion of the use of impact factors and the h-index in science, see Bornmann and Marx (2013a & b).
[8] We strongly recommend using scripting languages for automatic processing.

2 Challenges and Trial and Error

WOS calculates citation metrics on the basis of the set of papers it outputs which in turn depends on the individual search parameters that are entered. Stated differently, it is not possible to just enter the name of an author and get a complete overview of his or her scientific "impact" in the form of relevant citation parameters. Moreover, WOS has a strict journal admission policy regarding criteria such as the type and quality of the journals included in the database. Hence, we expect to obtain slightly fewer results of published papers in WOS as compared to the complete list of papers provided in the CV of the respective author. While this is an inaccuracy we have to accept, we strive to avoid the case of papers being included in the WOS output set that are not written by the respective author under study. The accurate selection of the papers under consideration is crucial to obtain undistorted citation metrics.

Our first approach is to search for an author's full last name and first initial using the WOS advanced search and then to extract the output of all indicated papers. We are reluctant to search for the full first name in addition to the full last name because this would filter out relevant papers which just carry the author's last name or the last name and the initial of the first name. However, by searching for the full last name and the first initial only the main problem is that the output of published papers cannot be unambiguously linked to a particular author under consideration. By comparing the WOS output with the CV of the respective author under study, we see that the WOS output contains papers from irrelevant authors with the same last name and the same first initial. Hence, we have to aim for a trade-off between minimizing the number of relevant papers that are excluded and excluding all irrelevant papers. The WOS menu provides refinements to rule out specific papers from the search results that help us to manage this trade-off. However, to exclude all irrelevant papers manually is extremely time-consuming. In the worst case, one would have to compare one by one the more than 10,000 published articles of the 488 authors under study listed in their CVs with the WOS output. Hence, we try to develop a different method that is both accurate and efficient.

We contact Dr. Hermann Schier (Information Service for the Institutes of the Chemical Physical Technical Section of the Max Planck Society), as his expertise with respect to data collection and creation using WOS and Scopus has been an extremely helpful source of information for us in previous research projects. In particular, we discuss whether there is an accurate and efficient method to obtain an output set of all published papers that matches perfectly with the authors under consideration. Although it is not possible to provide additional user rights in order to simplify our data collection within WOS, Mr. Schier suggests trying the Scopus citation database instead.

In fact, Scopus relates single authors with individual output sets of published papers by a search for just the full last name and the first initial. Hence, the Scopus database has the potential to significantly simplify our data collection. Nevertheless, after some intense trial and error with various samples, we decide not to use Scopus but WOS for two reasons.

First, the Scopus algorithm does not work with sufficient accuracy. Even though the Scopus search directly relates published papers to the authors under consideration, the output set of papers is far from complete (compared to the CV). In addition, the fact that the output contains some irrelevant papers suggests that the Scopus matching process does not yet work accurately. Hence, in order to ensure the consistency and integrity of our data creation process using Scopus we would again have to double-check the output set for every author manually by comparing the Scopus output with the available CV information. Second, the data converted into spreadsheets via Scopus is very hard to handle compared to WOS. It would be too time-consuming to rearrange and modify the data in spreadsheets in a way that would allow further analysis. Moreover, rearranging 488 spreadsheets would be rather prone to error.[9] In contrast, WOS provides spreadsheets that can be easily converted into the preferable form for subsequent evaluation.

In conclusion, we decide to primarily use WOS (complemented by eligible Scopus features for double-checks) and to search for the full last name and the initial of the first name. However, we also deem it necessary to refine the output according to a standardized procedure. We develop this procedure in order a) to guarantee that the data collection follows exactly the same process for every author and b) to accurately gather the required data in a time-efficient manner. More specifically, we determine specific refinement options in order to significantly reduce the likelihood of false output. We specify the exact steps of this process in the following.

3 Procedure of Data Collection

First, and prior to the actual search on WOS, we download the CV from the personal website of the researcher under study or the website of the institution the researcher is affiliated with. The CV is an important document in the data creation process for the following reasons. It provides an overview of the research areas of an author, which are the basis for our refinement strategy that we discuss below. It indicates the period of time in which the author has published articles. It provides exact information on potential middle names and the institution(s) an author is affiliated with.

Second, we search the author via Scopus. Typically, this immediately delivers accessible results since the author can be selected directly by his or her full name. Although Scopus has the above-mentioned drawbacks, its output serves as a benchmark for the total number of publications of the respective author under study retrieved from WOS. In addition, in cases where the author search on WOS does not lead to any results, a comparison with the Scopus output indicates whether the WOS search is too restrictive or the author has not yet

[9] Again, we strongly recommend using scripting languages for automatic processing.

published in WOS-listed journals.[10] After these initial steps, we perform the actual author search on WOS according to the following procedure.

3.1 Identification via Researcher ID

We search the Researcher IDs of the authors under study provided by Thomson Reuters.[11] The Researcher ID unambiguously identifies an author and allows WOS to accurately and directly output the corresponding list of publications without further refinements and adjustments. Stated differently, the identification of an author via Researcher ID is the best-case scenario. However, as of October 2013, the Researcher ID has several limitations. The major problem is that only a very small number of authors actually choose to have a Researcher ID (less than 8% in our sample). In addition, the WOS engine sometimes does not recognize the ID or shows no results despite the existence of the ID. In some rare cases, different authors than the author under investigation are listed by WOS when the ID is entered. Finally, one of the authors under study had two IDs, each delivering different results. In theory, the Research ID is a helpful tool. However, its practical utility is very limited due to the above-mentioned limitations. Hence, a more extensive search is required in the majority of the cases where a Researcher ID does not exist or where it does exist but delivers inadequate results.

3.2 Identification via Author Names and Refinement Options

We search for a particular author by inserting the full last name and first initial in the WOS advanced search.[12] We omit potential middle names in this first step. The results now show all registered authors with this name combination including those who are not relevant for our research. In extreme cases of common names such as Wang or Smith, the WOS output is a list of about 75,000 articles. By adopting the following procedure, we filter and refine those results to finally obtain only those papers published by the author under study.

First, we filter the results by name, which is done by the filter option *Author*. The author's CV or website gives information about potential middle names. Identification via middle names is an important (but not flawless) means to refine the WOS output. If a middle name exists, we refine the search results by full last name and first initial as well as by the full last name and first and middle initials. For instance, in the case of Robert James Smith, we only keep those results with Smith R and Smith RJ. However, if no middle name is indicated on the personal website or in the CV of the author, we may not simply conclude that an author does not have one (according to WOS). For example, it is possible that an author does not

[10] These double-checks reveal that in some cases the Scopus output is incorrect and that several authors with the same name are listed as one author.

[11] The Researcher ID can be obtained through an author search on www.researcherid.com (last visited October 15, 2013).

[12] By doing this, we ensure that the search results also include those papers that were tagged on WOS only with the last name or the last name and the first name initial.

use the middle name in some (more recent) publications, while in other (earlier) publications the middle name is still indicated. Hence, in the majority of the cases where a middle name does not exist or leads to ambiguous results, filtering only by name does not lead to accurate results and further refinements are required.

Second, WOS provides refinement by *Research Areas*. While this is a helpful device to filter the WOS output, an article often relates to more than one research area. For instance, WOS may categorize an article published in the field of behavioural economics into the research areas psychology, behavioural sciences and business and economics. In the majority of the cases, only selecting articles published in business and economics produces accurate results (as double-checks with CV information show) and allow us to exclude all non-economic articles. However, it also becomes apparent that some authors, especially those with a background in statistics, often have joint projects with researchers from other fields such as medicine and chemistry. In these cases, we select more than just the Research Area business and economics to increase the set of potentially relevant publications.

Third, an additional option to reduce the size of the search results is to set a threshold for the year of the first publication of the author under consideration. More specifically, we set the year before the author's first publication as the threshold to ensure that WOS outputs all articles of the author under study. This filter turns out to be particularly useful for young authors. In addition, it allows us to exclude authors from the search process when there are two authors with the same name who published in different periods of time.

Fourth, WOS allows filtering for institutions. This filter, together with the CV information on the institutions authors are affiliated with, allows us to further reduce the size of the search results. Note that the so-called *WOS Categories* provide another useful refinement option if the CV of the author under study clearly indicates his or her research field(s).

Fifth, we manually compare one by one the publications specified in the author's CV and the articles that WOS outputs to ensure that the WOS output only includes publications of the author under study. In some cases, it is necessary to manually deselect articles that WOS includes in its output but that are not listed in the author's CV. Of course, this can be time-intensive, particularly for productive researchers.[13] However, this last double-check turns out to be crucial to ensure the consistency and integrity of our data-creation process. Hence, we strongly recommend final double-checks in similar endeavours.[14]

Finally, as one major objective of the analysis in Andreoli-Versbach and Mueller-Langer (forthcoming) is to establish a connection between the author's quality, as measured by citation metrics, and data sharing, we create a citation report that can be directly extracted

[13] Nine of the authors under study have published more than 100 articles in WOS-listed journals.

[14] In general, it turns out to be useful to examine the results after each of the above-mentioned steps. In particular, the refinement by research area "Business and Economics" is often sufficient to retrieve articles of the respective author only. In this case, one can skip further refinements and directly compare the WOS output with the entries in the author's CV.

from the final WOS output set.[15] This report provides author-specific citation metrics based on the actual list of selected publications, e.g., total citations, average citations per paper, citations per year and h-index.

4 Conclusion

At first glance, the retrieval of author-specific citation metrics and publication data for a particular sample of 488 authors using "Web of Science" (WOS) and Scopus appears to be an extremely straightforward task. However, we experience several problems along the way that are inherent in the set-up of the databases and the way they refine and produce their output. To illustrate, WOS, without any further refinements, outputs a list of about 75,000 articles in the case of common last names like Wang or Smith. Hence, we extensively experiment with the refinement options provided by WOS to identify an accurate and efficient approach in order to ensure the integrity of our data creation process and provide the interested reader with a guideline for similar endeavors. The accuracy of the WOS output is particularly crucial for any analysis based on WOS citation metrics as WOS produces its author-specific citation metrics on a case-by-case basis depending on the actual list of selected publications.

In fact, WOS provides some useful refinement options that help to exclude irrelevant papers from the WOS output list. However, due to the various limitations of the search engines presented in this guide it turns out to be necessary for each author under study to double-check one by one the WOS list of publications with the publication information provided by the author's CV before generating the required citation metrics.

[15] This is done via the option *Create Citation Report* at the beginning of the WOS record list.

References

Andreoli-Versbach, Patrick and Frank Mueller-Langer, Open access to data: An ideal professed but not practised, Research Policy, forthcoming.

Bornmann, Lutz and Werner Marx (2013a), How good is research really?, European Molecular Biology Organization (EMBO) Reports, 14(3), 226-230.

Bornmann, Lutz and Werner Marx (2013b), How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations, forthcoming in: Scientometrics.

Furman, Jeffrey L. and Scott Stern (2011), Climbing atop the shoulders of giants: The impact of institutions on cumulative research, American Economic Review, 101(5), 1933-1963.

Gaulé, Patrick and Nicolas Maystre (2011), Getting cited: Does open access help?, Research Policy, 40(10), 1332-1338.

Hitt, Michael A. and Charles R. Greer (2012), The value of research and its evaluation in business schools: Killing the goose that laid the golden egg?, Journal of Management Inquiry, 21(2), 236-240.

McCabe, Mark J. (2002), Pricing and mergers: A portfolio approach, American Economic Review, 92(1), 259-269.