Scientometrics (2018) 117:705–720
https://doi.org/10.1007/s11192-018-2895-3

Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates

Valeria Aman[1]

Received: 5 October 2017 / Published online: 1 September 2018

Abstract

In this paper I outline how author identifiers enable to track international mobility of scientists. Authorship systems help to distinguish among similar names and provide information on affiliations and thus countries of stay. This study explores the relation between CV data and Scopus data in regard to tracking international mobility of scientists. To test the consistency and applicability of data on mobility episodes, residence countries as provided in CVs of a set of German scientists were compared against country information in the affiliations of their publications. Therefore, the CVs of Leibniz laureates were coded for the period 1996–2015 and their publications were gathered on the basis of Scopus author ID. Results show that the majority of scientists under study have a single author ID (68.4%). However, there are laureates with so-called 'split identities' where more than one author ID exists. Most of them have a dominant author ID that covers the majority of their publications and one or more additional IDs with only a few publications causing these split identities. Recall statistic shows that the use of the dominant author ID of each laureate would result in around 97% of their publication output. In contrast, the precision of Scopus author ID proves to be high. A random sample shows that all publications assigned to a specific author ID relate to a single individual, so that the precision statistic would yield 100%. Further results show that the registry systems ORCID and ResearcherID are no alternatives to Scopus author ID, because a minority of laureates make use of these identifier systems and data is often incomplete. Unlike ORCID and ResearcherID that suffer from a selection bias as those scientists who remain in science maintain their author profiles, Scopus author ID exists for every author publishing in sources covered by Scopus. The comparison of mobility data in Scopus versus CV data shows that bibliometric data is suitable to identify a scientist's international mobility and appears to be a good solution if there are no CVs available or if they are incomplete. Furthermore, the reasons for inconsistencies in mobility data are discussed. These mainly reside in the lack of co-author affiliations, incomplete CV data, and other minor reasons.

Keywords Scientific international mobility · CV data · Bibliometric data · Author identifier system · Scopus author ID · ResearcherID · ORCID · Leibniz laureates · Data consistency

Extended author information available on the last page of the article

Introduction

Despite the increased interest in studies of scientific international mobility, scholars are often confronted with insufficient data on mobility flows (Pirralha et al. 2009; Børing et al. 2015). Statistical data on the mobility of scientists can be obtained with a diversity of methods. Apart from the direct collection of information via surveys or interviews, CVs and bibliometric data are a valuable source for studies of scientific mobility. Because mobility episodes take place along a scientist's career, the analysis of scientists' trajectories via CVs was established as a promising method (Cañibano et al. 2008; Woolley and Turpin 2009). By studying affiliations of publications over time, bibliometric data is equally suitable to indicate the movement of scientists from organisation to organisation and thus from country to country. Laudel (2003) was the first to combine bibliometric data with CV data to test their usability of studying scientific mobility. The comparison of three bibliometric databases (Web of Science, PubMed, and INSPEC) with CV data showed that bibliometric data reflected the mobility episodes of scientists with a time lag of 1 year.

Working with bibliometric data one major difficulty is the identification of the full scientific output of an author. Many different scientists share the same name and at the same time names and affiliations may change over time. Another problem is that many authors only list their initials instead of their full first names (Lerchenmueller and Sorenson 2016). One can overcome these difficulties by relying on manual search of CVs gathered from the Internet. This approach ensures a high level of accuracy but is time-consuming and therefore limited to small populations.

Author identification systems enable to deal with these problems of *homonyms*, i.e. different authors sharing the same name and synonyms, i.e. an author having more than one name (see Moed and Halevi 2014). Author IDs can be used to trace individuals over time and to explore the international mobiltiy of scientists. The author identification system in focus of this study is the Scopus author ID implemented in *Elsevier's* Scopus database. Scopus assigns each document an author ID via an algorithm that matches authorship on several criteria such as affiliation, subject area, source titles, or co-authors (Moed et al. 2013; Moed and Halevi 2014). Whereas some studies reported the Scopus author ID to be precise (Kawashima and Tomizawa 2015), other studies found it acceptable to trace scientists' mobility if results on the macro-level are adjusted to bypass potential flaws (Moed and Halevi 2014; Conchi and Michels 2014). The study by Kawashima and Tomizawa (2015) in which the precision and recall of Scopus author ID is estimated by being matched to the Japanese funding database KAKEN has parallels to the one presented here.

Another author ID system is offered by *Clarivate Analytics*' Web of Science database hosted on the Web of Knowledge platform. This so-called ResearcherID was introduced in 2008 as an author identification system. Unlike the Scopus author ID that is automatically assigned to all publications in the database, authors have to register for a ResearcherID. This enables authors to include information about their affiliations, add alternative spellings of their name, identify which publications are theirs (and which not), add subjects and view their citation metrics.

The pressure for a single author identifier led to the founding of the Open Researcher and Contributor ID (ORCID) initiative. Unlike Scopus author ID and ResearcherID, ORCID is run by a non-for-profit organisation with a rotating board and more than 600 member organisations (Carter and Blanford 2017). Authors can register and receive an identifier which can be used whenever work is submitted for publication. In 2012, when the ORCID registry was launched, *Clarivate Analytics*' predecessor *Thomson Reuters* announced it would incorporate ORCID identifiers into its scientific offerings to ensure a bi-directional link between ORCID and ResearcherID. Likewise, the registration for an ORCID allows importing publications from ResearcherID as well as from Scopus author profiles.

In the centre of this case study is the Scopus author ID that is available for all authors in Scopus. ResearcherID and ORCID are used to validate the precision of Scopus author ID and to identify the extent to which authors under study have registered for ResearcherID and ORCID, respectively. So far, there are only a few methodological discussions of the use of Scopus author ID on a micro-level to track international mobility. Therefore, this study compares the quality of data on scientists' international mobility derived from CVs to data derived from bibliometric databases. A special focus is on the identification of scientists in Scopus and the precision of affiliation data. To answer the research question of whether Scopus author ID suffices to study international mobility, I focus on a group of individual scientists. Whereas in some countries, databases have been created with standardised scientists' CVs (LATTES in Brazil, DeGóis in Portugal, Europass in Europe), no such comprehensive database exists for German scientists. Therefore, I focus on Leibniz laureates whose online-published CV is an integral part of the awarding process. The Leibniz Prize is the most important research prize in Germany, annually awarded by the DFG (*Deutsche Forschungsgemeinschaft—German Research Foundation*). It is the highest endowed research prize in Germany with a maximum of €2.5 million per award to improve the working conditions of outstanding scientists and to expand their research opportunities. On the basis of these laureates, the goal is to study the usefulness of CV data and bibliometric data for tracking the mobility of scientists.

CV data were coded according to the countries laureates moved to and the duration of their stay within the 20-year period ranging from 1996 to 2015. To ease the comparison of data, a bibliometric data set of the laureates was created on the basis of their publications in Scopus. These two methodological approaches to study international mobility are discussed in the following. Methodological failures and how to cope with them are addressed to finally assess the applicability of bibliometric data for studying scientific international mobility.

Not only is little known about methodological aspects of tracking international scientific mobility, but also about how scientists move, where they move and the importance of international mobility for their careers (Sugimoto et al. 2017). Moreover, our knowledge of eminent scientists and especially their mobility patterns is still very limited.

Data and methods

Data base

The data were retrieved from *Elsevier's* Scopus database that is licensed as custom data at the *Competence Centre for Bibliometrics*.[1] Scopus offers an automatic author ID that combines all publications of an author under a single ID to handle common first and last names (Moed et al. 2013). In order to associate authors with their publication oeuvres, the algorithm uses a multifaceted approach where name spelling variants, affiliations, co-authors, subject areas and the prior publication history are taken into account. The algorithm aims at higher levels of precision than recall. Thus, as soon as a publication cannot

[1] Competence Centre for Bibliometrics. http://www.forschungsinfo.de/Bibliometrie/en/index.php?id=home.

be assigned to an existing author ID, a new profile with a new author ID will be created under which the publication appears (Moed and Halevi 2014). The algorithm is supplemented by an author feedback system where authors can indicate whether publications are wrongly attributed to their profiles.

In addition to the Scopus database, data were retrieved from *Clarivate Analytics'* Web of Science database that is also licensed as custom data at the *Competence Centre for Bibliometrics*. This database includes the ResearcherID as well as ORCID for those authors who have signed up for either of these author identifier systems.

Data selection

The data set encompasses all 193 laureates who were awarded a Leibniz Prize between 1999 and 2016. CVs of laureates being awarded a prize between 1986 and 1998 are not available online in the DFG archive. The CVs of all 193 laureates were downloaded from the Internet. In case it was not possible to find the most up-to-date CV, the CV from the DFG homepage was used. CVs were coded according to the country of residence and the duration of stay. Only those residence countries were retrieved from CVs where it was evident that a laureate has stayed there for at least 1 month. Information on memberships that are not specified by a year or duration was not coded, because it does not provide precise information on a physical stay abroad.

To identify author IDs, laureates were searched in a licensed database version of Scopus. The search string to be matched against all authors in the database consists of the last name and the initial of the first name of a laureate, e.g. 'bradke, f' for the laureate Frank Bradke. As a result, the following data is retrieved: the indexed name Frank, B., the last name Bradke, the first name Frank and an e-mail address if it is attached to a publication. The identification of author IDs that truly belong to a laureate in question was done manually. Thus, whenever there is more than one author matched to an indexed name, the publications attached to a Scopus author ID were randomly checked. Therefore, randomly chosen articles assigned to an author ID were looked-up on the Internet. The affiliation on the paper and the comparison with organisations the laureate has been affiliated to according to CV data revealed whether a specific author ID really belongs to a laureate or not. Whenever possible the institution information that is often part of e-mail addresses was compared against the institutions at which the laureate has stayed according to CV data. This identification procedure guarantees that all Scopus author IDs belonging to a single laureate are found. Thus, the problem of 'split identities' can be manually handled.

After the laureates were identified, their Scopus publications were collected for the 20-year period (1996–2015). Only journal articles and reviews are considered. To tackle the problem of 'merged identities' a sampling was applied to prove whether a single author ID of a laureate combines other publications than those of the laureate. To this end, 10% of publications of 10% of author IDs were randomly selected and manually proved. The original publications were checked in terms of the first and last name and the affiliation. On the basis of this sample the author ID yielded no erroneous attributions, so that there is no evidence that publications are mistakenly attributed to author IDs. Based on this data, I evaluate the recall and precision of the Scopus author ID that is presented in the result section.

In terms of the author identifiers implemented in WoS, all laureates were manually searched on the basis of their last name and the initial of their first name. Similar to the approach in the Scopus database, publications (articles and reviews only) of laureates with a ResearcherID and ORCID respectively, were collected for the same time period.

Methods

To evaluate Scopus author ID recall and precision were estimated. It is not possible to provide an overall recall that indicates how many of the publications penned by one and the same laureate can be found in Scopus under a specific author ID. This is because there are no complete publication lists of laureates. There are only a few CVs that can be found on the Internet including publication data, most often only a selection of recent or highly-cited publications. However, I use the term 'recall' to refer to the amount of papers that can be found under a dominant author ID, thus an author ID that subsumes the majority of publications of the author in question. I define 'precision' as the share of publications that truly belong to an author as represented by an author ID.

Another method used to validate the accuracy of Scopus author ID can be described as follows. All publications assigned to ResearcherID or ORCID in WoS were matched with publications in Scopus according to the same publication year and source title, whereas the article title does not have to be identical, because it can vary in Scopus and WoS due to different approaches to transliterate formulas or to handle characters. Therefore, the edit distance (Levenshtein algorithm) was used to allow a valid match to vary up to 10% in the characters (normalized by the length of the article titles). This method allows retrieving publications in Scopus that are assigned to author IDs other than those that were initially identified.

Not every laureate has a continuous publication activity and thus some laureates lack country information for some years. Missing country information can be filled in two different ways. The first assumption is that if authors have stayed in the same country of their last known residence, the missing country information of a year can be filled with the information from the previous year. This can be done if at a later point, country information is available. This procedure is termed in the following 'forward filling' and was also applied by Conchi and Michels (2014).

The second assumption is that if an author has country information according to a publication in 1 year but no country information of the previous year, the missing information is filled with the publication information. In this case the most recent publication denotes the end of publication activity. This procedure is termed 'backward filling' and appears to be more appropriate if one assumes that a publication results from a previous stay in a country. The two procedures are exemplified in Table 1. The automatically-filled country information is in italics, whereas the original information from the publications in Scopus is in bold print.

In many cases authors have resided and published from different countries in 1 year. Then these countries are weighted, so that if an author published from two countries in 1 year, each country is counted one half.

The main methodological question is how high the accordance between mobility episodes from Scopus publication data versus CV residence data is. One main reason for potential discrepancies is assumed in multiple affiliations of authors. Therefore, it is tested how the exclusion of multiple-affiliations publications influences the precision of results.

Table 1 Example of forward filling (FF) and backward filling (BF) of author's missing country information in Scopus

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
FF		**DEU**	*DEU*	**USA**	*USA*	*USA*	**DEU**	**DEU**	*DEU*	*DEU*	**GBR**	*GBR*	*GBR*	*GBR*
BF	*USA*	*USA*	**USA**	**DEU**	*DEU*	**DEU**	*NLD*	**NLD**	*DEU*	*DEU*	**DEU**	**DEU**	**USA**	

Results

To start with, I provide some background information on the laureates under study. There are 193 laureates who were awarded a Leibniz Prize between 1999 and 2016, of which 161 are male (83.4%) and 32 female (16.6%). Their average age at the time of award is 47.7 (standard deviation: 5.2). Taking the OECD Frascati classification of science and technology (FOS)[2] as a basis, we can derive from Table 2 that the majority of laureates belong to the Natural Sciences.

Figure 1 informs about the up-to-dateness of CV data used. The latest available CV on the Internet was searched to have a complete view on the trajectory of laureates. The histogram shows that the majority of CVs are from recent years. However, there are still some CVs that are out of date and are based on the year of prize award. The reason is that the year of award coincides with the end of the scientific career of a laureate.

Table 3 provides an overview of the number of Scopus author IDs per laureate. In most cases all publications of a laureate are combined under a single ID (68.4%). However, there are 56 split identities where more than one author ID was found for one and the same author. Most often one profile contains the majority of publications and several smaller profiles only 1–3 papers (Moed and Halevi 2014).

A more detailed analysis shows that of those 38 split identities with two different author IDs, 25 have an author ID that subsumes at least 95% of their publication output, and 5 split identities who have an author ID subsuming between 90 and 95%.

There are two special cases where in one and the same publication two different IDs were assigned to one individual. These two laureates have one author ID that includes 100% of their publications and an additional ID resulting from a single publication that is characterized by a high number of co-authors. For those 32 laureates with a dominant author ID comprising more than 90% of their oeuvre it would suffice to work with this major ID. The publications of the other 6 split identities are distributed more evenly over the two author IDs. Thus, it would be important to identify both author IDs to have a complete overview of their publication output.

Overall, the recall statistic can be expressed in the following way: Taking only the dominant author ID for each laureate, one would find on average 97.14% of their publication output. The other three percent of publications are assigned to additional author IDs causing split identities. The share of papers assigned to a dominant author ID ranges from 7.18 to 100%, the median being 100% and the standard deviation 9.93%. Random sampling of publications and their manual check on the Internet as described in the method section shows that all looked-up publications are correctly assigned to an author ID, thus the precision of the sample is 100%.

Split identities are mainly caused if no e-mail address exists, which is obviously an important matching criterion for the Scopus author algorithm. Detailed analyses show that another major reason causing split identities are recent publications from the years 2014 and 2015 (which may differ in their topic, journal used or co-authors).

Of all 42 laureates who have a dominant author ID comprising at least 90% of their total publication output, 18 laureates have one or several additional author IDs that subsume publications only from the years 2014 and 2015. Regarding this issue, *Elsevier* states that the algorithm rather creates a new author ID than taking the risk to subsume new publications under an existing ID. According to *Elsevier*, these split identities may be gathered

[2] OECD (2007). REVISED FIELD OF SCIENCE AND TECHNOLOGY (FOS) CLASSIFICATION IN THE FRASCATI MANUAL, 26 Feb 2007. http://www.oecd.org/sti/inno/38235147.pdf.

Table 2 Overview of disciplines the 193 laureates under study belong to

OECD_FIELD	Number of laureates	Share in %
Natural sciences	107	55.4
Humanities	29	15.0
Medical and health sciences	25	13.0
Engineering and technology	17	8.8
Social sciences	15	7.8
	193	100

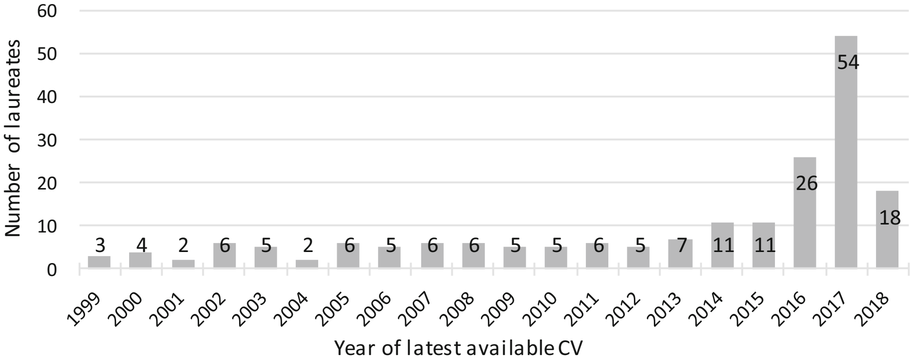

Fig. 1 Overview of the up-to-dateness of CVs of laureates

Table 3 Overview of the number of author ID matched for a laureate and share in %

Number of author IDs	Number of laureates	Share in %
0	5	2.6
1	132	68.4
2	38	19.7
3	9	4.7
4	4	2.1
5	2	1.0
6	1	0.5
11	1	0.5
143	1	0.5
	193	100

in future under a dominant author ID when a critical mass of publications is reached that proves that they belong to an author with an existing author ID. *Elsevier* also relies on the authors themselves who can make sure that their recent publications are covered under a pre-existing ID and not under a new ID.

However, it is unclear how fast and how precisely Scopus resolves these split identities and in how far *Elsevier* relies on authors to fix inaccuracies concerning their IDs. There is no information on how many authors are aware of their Scopus author ID and keep it updated but *Elsevier* has established a user-friendly *Support Center* where authors learn how to correct their author profile.

Detailed analyses also show that international mobility and new affiliations cause split identities. The group of laureates with split identities has an average affiliation count of 1.24 whereas laureates with a single author ID have an average affiliation count of 1.16. As Kawashima and Tomizawa (2015) state, it is difficult to distinguish an author who was internationally mobile and changed affiliation and co-authors from two different authors with the same name and different co-authors and affiliations (p.1070). Moreover, there is high probability that authors retain a footing in their country of scientific origin throughout their career and accumulate more foreign affiliations with every international stay abroad (Sugimoto et al. 2017, p.30).

The publication productivity may also influence split identities. Those laureates with split identities have an average publication count of 176.0 whereas authors with a single author ID have on average 112.9 publications. It is only natural that the more productive an author is, the higher the probability is that the topics of the publications differ, the sources used just as the co-authors. All these criteria challenge the Scopus algorithm in the identification process.

As Table 3 shows, there is an extreme case where 143 IDs were found for a single laureate. On the one hand, every publication was assigned a different author ID (0.48%) in an ascending order, which is most probably a Scopus mistake. On the other hand, the laureate in question has additional author IDs combining up to 7.18% of all his publications.

A minority of 5 laureates (2.6%) is not represented by publications in Scopus and therefore has no author ID. They belong to the social sciences and humanities. The following analyses thus build upon 188 laureates for which both CV and Scopus data are available. This group of scientists is characterized by a high publication output. Overall, they have published 24,756 journal articles and journal reviews between 1996 and 2015 according to Scopus. The distribution of publications over laureates is skewed as can be derived from Fig. 2. The laureates are in ascending order of their publication number. On average, each laureate has published 131.7 papers between 1996 and 2015, the standard deviation is 113.9.

Due to problems resulting from double affiliations, the country information of laureates is also determined on the basis of single-affiliation papers (23,524; 95.02%).

Unlike Scopus author ID, ResearcherID and ORCID are opt-in systems and only include those who have registered with ResearcherID or ORCID, either by themselves or by an institutional manager. As Table 4 shows, a total of 48 laureates have a ResearcherID, whereas the number of laureates with an ORCID is 40 (24.9% and 20.7%, respectively of

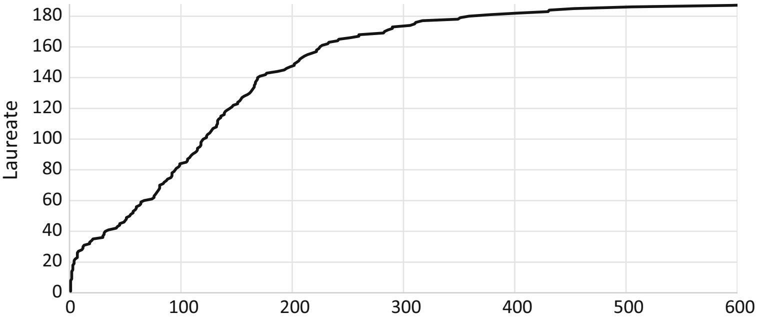

Fig. 2 Cumulative distribution function of publications per laureate in 1996–2015

Table 4 Comparison of ResearcherID and ORCID data in WoS with data in Scopus

	ResearcherID	ORCID
Number of laureates	48	40
Publication count in WoS	7104	4711
Publication count in Scopus	9337	7855
Share of publications in WoS	76.08%	59.97%

overall 193 laureates). Only 31 laureates have both a ResearcherID as well as an ORCID (16.1%).

The table shows that the number of publications of those 48 laureates who have a ResearcherID is much smaller than the number of publications covered in Scopus, which is mainly due to the broader coverage of sources in the Scopus database. The match of article titles in WoS and Scopus revealed some inconsistencies in dealing with publication and document types, e.g. publications in Scopus that are attributed to the document type ‘Review’ are in WoS declared as ‘Software Review’.

Furthermore, the results show that the publication counts of laureates vary among the three author IDs. Out of 31 laureates with both a ResearcherID and an ORCID, only 20 have the same publication number. Moreover, it is striking that some authors use ORCID infrequently (e.g. whenever it is required by a publisher) so that the number of publications assigned to ORCID is lower than that to ResearcherID of one and the same laureate (see Table 4). Thus, whereas some authors have only one of their articles assigned to a ResearcherID or an ORCID, others seem to have a complete list of publications that is also in accordance with data in Scopus.

A major advantage of the analysis of ResearcherID and ORCID is that it allows revealing those Scopus author ID that were missed in the manual search for author IDs. Therefore, publications of laureates in WoS assigned to either ResearcherID or ORCID were matched in Scopus on the basis of their article title, author last name and source title as well as publication year. Hereby, articles were identified in Scopus by those laureates who have an additional Scopus author ID that was not identified in the initial manual search for author IDs. However, these additional IDs comprise a small set of publications (1–3) and the majority is still gathered under a dominant author ID that was identified as such. According to ORCID, there are three authors (with 2 publications each) who have an additional Scopus author ID that was not found in the initial manual search. ResearcherID reveals that the same three author IDs were missed in the manual search plus another laureate with only one author ID and publication. The reason why these additional author IDs were not found are spelling variants, e.g. *Weige* instead of *Weigel, Mattias* instead of *Matthias* or *Dan* instead of *Daniel* as in the majority of previous publications. Optical character recognition can also lead to spelling mistakes, e.g. the first name ‘Ilme’ was recognized as ‘Urne’ and was therefore not identified in the initial search because it starts with another letter. These four Scopus author IDs missed show that it is not the algorithm that causes split identities but often the authors and editors who pay little attention to the spelling of their names and the inclusion of e-mail addresses. An e-mail address attached to a misspelled author name might help the algorithm to assign the publication affected to a dominant author ID.

Country information

The following figure illustrates the number of countries laureates are affiliated with, according to CV and publication data. The publication data was distinguished into country information on overall publication data and country information on the basis of publications with single affiliations.

It can be derived from Fig. 3 that according to CV data, 68 laureates were non-mobile and stayed in Germany between 1996 and 2015. A majority of 86 laureates was in another country than Germany and the rest in two or more countries other than Germany. The publication data in Scopus suggest that the laureates under study published from more countries than those in which they resided. This discrepancy is mainly ascribed to double affiliation papers. The limitation to single-affiliation papers produces results that approach those derived from CV data. On the whole, Leibniz laureates seem to constitute a highly internationally mobile group of scientists. In contrast, a broad analysis comprising 16 million individuals publishing between 2008 and 2015 showed that only 4% were internationally mobile (Sugimoto et al. 2017, p.29).

To compare residence country information with those that can be derived from publication data, it is important to be aware of the publication continuity. The following Fig. 4 provides an overview of 188 laureates and the number of years they are represented by publications in Scopus. A majority of 97 laureates has published throughout the years 1996–2015.

For those laureates who have missing country-year information, forward and backward filling was applied. In addition, the filling procedures were tested for all publications and only those with single affiliations. The results are provided in Table 5. The average share of residence country years missing in the forward-filled data on the basis of all publications is 14.8%, whereas the average share of forward filled country year combinations not in accordance with CV data is 12.4%.

The best result is produced with backward filling and single-affiliation papers. The average share of residence-country years missing in the backward-filled data on the basis of single-affiliation papers is 12.3%, whereas the average share of backward-filled country

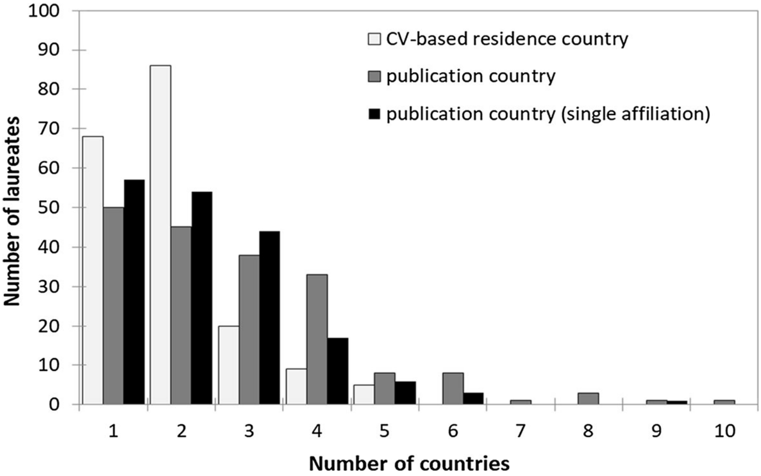

Fig. 3 Number of countries of laureates between 1996 and 2015 according to source

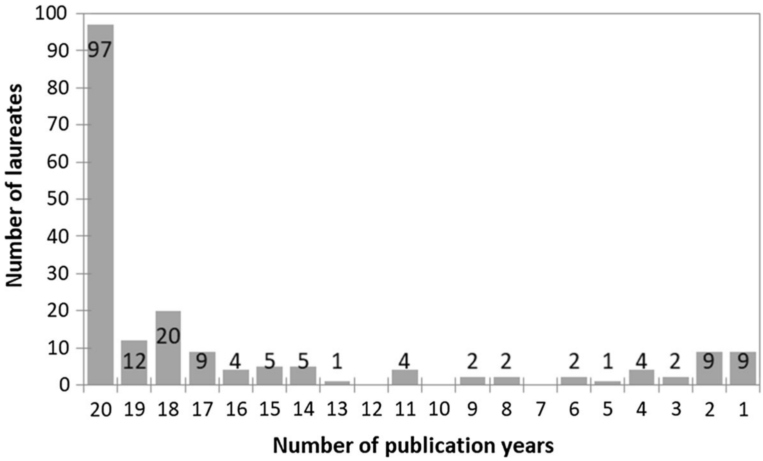

Fig. 4 Publication activity of laureates according to Scopus data

Table 5 Overview of the accordance of CV and bibliometric data, distinguished by affiliation and filling procedure

Affiliation type	Filling procedure	Residence-country years missing in %	Filled country year combinations not in accordance with CV in %
Multiple	None	20.7	13.0
	FF	14.8	12.4
	BF	10.5	13.3
Single	None	25.2	8.4
	FF	16.8	8.4
	BF	12.3	9.3

year combinations not in accordance with CV data is 9.3%. The information on residence-country years missing can thus be reduced by half from 25.2 to 12.3% with the backward-filling procedure.

The discrepancy of country information as illustrated in Fig. 3 was analyzed in more detail. Those publications were gathered that have country information that is not in accordance with CV data. There are overall 363 (1.47%) publications in Scopus that provide country information that does not occur in CV data. Overall, 113 laureates are affected by this inconsistency. Assuming that the inconsistency resides in multiple-affiliation papers, only those with single affiliations were considered. Then the publication number with inconsistent CV country information and publication country information diminishes to 209 (0.84%) and the number of laureates affected to 91. These 209 publications were manually checked to determine reasons for inconsistency. Table 6 comprises the reasons and the number of publications affected.

The table shows that the comparison of country information from CV and bibliometric data can reveal valuable information. The juxtaposition of these two sources of information on mobility data shows that there are two merged identities detected. These were not excluded because the majority of publications combined under the merged author ID are those of the laureates. The erroneously attributed publications of authors with the same last

Table 6 Overview of reasons for inconsistency between CV country information and Scopus country information

Reason for mistakes	No. of publications	Share in %
Erroneously attributed affiliation by publisher	2	1.0
Stay abroad was prior to 1996	5	2.4
Paper was published after the latest available CV	8	3.8
Stay is not explicitly mentioned in the CV	14	6.7
Merged identity	15	7.2
Country code problems in Scopus	17	8.1
Only information on the first author of the paper (web search)	20	9.6
Affiliation of the first author or any co-author was taken	128	61.2
Total	209	100

and first name apparently caused country information that is not in accordance with CV data. The most serious error that is responsible for inconsistent country information is caused in Scopus when affiliations of co-authored publications are mixed up so that the author in question receives the affiliation of any co-author of the paper.

Discussion and conclusion

This study has attempted to explore the strengths and weaknesses of CV data and Scopus data in regard to tracking international mobility. On the basis of German laureates whose residence countries (as provided in their CVs) were compared against the country codes in the affiliations of their publications, I analyzed the consistency of the data on mobility episodes.

The comparison of these two data sources shows that bibliometric data is suitable to identify a scientist's international mobility and appears to be a good solution if there are no CVs available or if they are incomplete.

However, CV data as well as bibliometric data have several shortcomings that prevent a straightforward application. The use of CV data requires high effort in regard to coding mobility episodes, whereas the use of bibliometric databases can become an elaborate task, especially when author data has to be purified from homonyms and synonyms. Scopus author ID has the potential to offer correct (precise) and complete (recall) lists of scientists' publications in the Scopus database. The active participation of authors identifying their own publications makes Scopus author ID an easily accessible disambiguation system of authorship for a large number of authors. However, results show that articles by a single individual are often assigned to more than one author ID causing split identities. This finding is consistent with Moed et al. (2013) who emphasize that the system puts more weight on precision than on recall. On the contrary, the evaluation of the precision of the algorithm shows that those publications that are assigned to an author ID truly belong to the author.

It is worth emphasizing that the Scopus algorithm is especially challenged for recent publication years. The algorithm rather creates a new ID than subsuming publications under an existing ID if incoming publications cannot be assigned unambiguously to an author ID. It remains unclear how fast and how precisely Scopus resolves these inaccuracies. Therefore, future research could investigate how fast and how accurately Scopus solves the issue of split identities. It is assumed that the author IDs are checked on a yearly basis. It is unknown though, to what extent *Elsevier* relies on authors to fix inaccuracies. However, this approach is influenced by a selection bias as those authors update their information who are active in science.

Although the Scopus author ID algorithm proves to be valuable, split and merged author profiles due to synonyms and homonyms remain a problem. While the approach to manually identify split author profiles proved feasible for a small sample, it would be problematic at a large scale as scientists must be identified unambiguously. The evaluation of the sample did not yield any merged identities. However, the detailed analysis of publications with inconsistent country information (in regard to CV data) revealed two merged identities. The publications by the authors who were erroneously merged with the laureates make up a small share in comparison to the laureates and were therefore not excluded.

Different national traditions of name attribution, incorrectly abbreviated, misspelled and misplaced first and last names affect all authors in science. Authors and editors are equally responsible for erroneous records in author profiles. Author name disambiguation is thus important for bibliometric analyses and remains a problematic issue when applied to individual authors (Smalheiser and Torvik 2009). However, Scopus author ID provides a neat solution to the author ambiguity problem within the scholarly research community. To avoid author misidentification, scientists can sign up to check whether publications are correctly attributed to their profile. The profile page is available for all authors in Scopus and does not require individual scholars to enter data. If the algorithm fails to identify all variations of an author's name, a request in the author feedback system can be sent to merge publications into a single profile.

The use of self-reported registries such as ORCID or ResearcherID may promise a higher level of accuracy. However, the results show that a minority of scientists under study make use of these systems and both alternatives raise concerns as to what form of selection bias might determine who signs up and organizes their entries in these self-reported registries (Lerchenmueller and Sorenson 2016). Although ResearcherID is available for almost a decade, the majority of scientists have not taken advantage of the ResearcherID capabilities or even set up a profile.

More and more journal editors require authors to use the ORCID system in the submission process and connect publications of an author to a single ID, e.g. those published by Springer or PLOS (Leopold 2016). Prior to the submission process, corresponding authors must register for ORCID and the employment and publication section is henceforth publicly available. Some journals even require all authors of a manuscript to sign up for ORCID before a paper can be published (Carter and Blanford 2017). Similar to plagiarism checking, ORCID is part of the commitment to ethical publishing (Carter and Blanford 2017). However, some journals such as CORR refrain from the use of ORCID, because it does not reliably identify that an author is who she or he claims to be (Leopold 2016). Thus, Scopus offers a true alternative to track international mobility due to the hybrid approach of assigning author ID to every author in the database and the option to keep track of publication portfolio behind an author ID.

These three author identifier systems can make it tedious for scientists to maintain all profiles. However, all author ID systems enable institutions to establish an administrator to upload a scientists' information, edit profiles and create publication lists. Especially for organisations, the accuracy of these claimed affiliations are vital for their understanding of their research portfolio, their research activity and their reputation and impact. The claim that a certain researcher is affiliated with a certain organisation is a fundamental component of research information.

Some limitations remain inherent to bibliometric studies of international mobility. Bibliometric research allows tracking mobility only to the extent that scientists publish and that the affiliation is stated on the publication and is linked to them. Moreover, a stay abroad does not necessarily result in a publication, and thus cannot be measured as such. CVs may provide information on stays abroad that do not necessarily result in publications that carry the affiliation of the host institution. Also, the publication delay has to be kept in mind, thus a mobility episodes that becomes evident through publications may result from earlier stays abroad. Finally, the coverage of journals in Scopus reflects only one aspect of scientific activity and can be insufficient for detecting mobility episodes.

As the results have shown, not all affiliations of co-authors of a publication are available, so that co-authors may receive the same affiliation as the first author in Scopus. Distortions may also result for scientists who move to another country but continue to publish with colleagues from their former institutions and from double affiliations, thus when scientists are simultaneously appointed at institutions that are located in different countries. Kawashima and Tomizawa (2015) therefore assume that Scopus author ID is more reliable for authors with low international mobility compared with authors with high international mobility and changing affiliations (p.1070).

Finally, the results presented in this paper are based on a rather small sample of eminent scientists. An analysis could be performed for 'average' scientists to see whether and how their mobility patterns affect the results and differ from the inconsistencies found for eminent scientists.

Acknowledgements The study was funded by the Bundesministerium für Bildung und Forschung (BMBF) under the Grant Number 01PQ16002. The data builds on the bibliometric database provided by the Competence Centre for Bibliometrics (Grant Number: 01PQ17001). I would like to offer my special thanks to Nicolai Netz, two student assistants and anonymous reviewers.

References

Børing, P., Flanagan, K., Gagliardi, D., Kaloudis, A., & Karakasidou, A. (2015). International mobility: Findings from a survey of researchers in the EU. *Science and Public Policy, 42*(6), 811–826. https://doi.org/10.1093/scipol/scv006.

Cañibano, C., Otamendi, J., & Andújar, I. (2008). Measuring and assessing researcher mobility from CV analysis: The case of the Ramón y Cajal programme in Spain. *Research Evaluation, 17*(1), 17–31. https://doi.org/10.3152/095820208X292797.

Carter, C. B., & Blanford, C. F. (2017). All authors must now supply ORCID identifiers. *Journal of Materials Science, 52*(11), 6147–6149. https://doi.org/10.1007/s10853-017-0919-7.

Conchi, S., & Michels, C. (2014). Scientific mobility: An analysis of Germany, Austria, France and Great Britain. Fraunhofer ISI discussion papers innovation systems and policy analysis. Abgerufen von. http://hdl.handle.net/10419/94371.

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061–1071. https://doi.org/10.1007/s11192-015-1580-z.

Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? *Scientometrics, 57*(2), 215–237. https://doi.org/10.1023/A:1024137718393.

Leopold, S. S. (2016). Editorial: ORCID is a wonderful (but not required) tool for authors. *Clinical Orthopaedics and Rlated Research, 474*(5), 1083–1085. https://doi.org/10.1007/s11999-016-4760-0.

Lerchenmueller, M. J., & Sorenson, O. (2016). Author disambiguation in PubMed: Evidence on the precision and recall of author-ity among NIH-funded scientists. *PLoS ONE, 11*(7), e0158731. https://doi.org/10.1371/journal.pone.0158731.

Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics, 94,* 929–942. https://doi.org/10.1007/s11192-012-0783-9.

Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics, 101,* 1987–2001. https://doi.org/10.1007/s11192-014-1307-6.

Pirralha, A., Fontes, M., & Assis, J. (2009). Assessing scientific mobility dynamics and impact: Drawing on the potential of electronic CV databases. In *Gehalten auf der ESA2009—9th conference of European Sociological Association*.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1–43.

Sugimoto, C. R., Robinson-Garcia, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature News, 550*(7674), 29. https://doi.org/10.1038/550029a.

Woolley, R., & Turpin, T. (2009). CV analysis as a complementary methodological approach: investigating the mobility of Australian scientists. *Research Evaluation, 18*(2), 143–151. https://doi.org/10.3152/095820209X441808.

Affiliations

Valeria Aman[1]

✉ Valeria Aman
aman@dzhw.eu

[1] German Centre for Higher Education Research and Science Studies (DZHW), Schuetzenstrasse 6a, 10117 Berlin, Germany