an open access journal

Check for updates

Citation: Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020).Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. https://doi.org/10.1162/qss_a_00031

DOI:
https://doi.org/10.1162/qss_a_00031

Supporting Information:
https://www.mitpressjournals.org/doi/suppl/10.1162/qss_a_00031

Received: 29 August 2019
Accepted: 14 January 2020

Corresponding Author:
Chun-Kai (Karl) Huang
karl.huang@curtin.edu.au

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

Comparison of bibliographic data sources: Implications for the robustness of university rankings

Chun-Kai (Karl) Huang, Cameron Neylon, Chloe Brookes-Kenworthy, Richard Hosking, Lucy Montgomery, Katie Wilson, and Alkim Ozaygen

Centre for Culture and Technology, Curtin University, Bentley 6102, Western Australia

Keywords: bibliographic data, data quality, open access, OpenCitations, research evaluation, university ranking, Unpaywall

ABSTRACT

Universities are increasingly evaluated on the basis of their outputs. These are often converted to simple and contested rankings with substantial implications for recruitment, income, and perceived prestige. Such evaluation usually relies on a single data source to define the set of outputs for a university. However, few studies have explored differences across data sources and their implications for metrics and rankings at the institutional scale. We address this gap by performing detailed bibliographic comparisons between Web of Science (WoS), Scopus, and Microsoft Academic (MSA) at the institutional level and supplement this with a manual analysis of 15 universities. We further construct two simple rankings based on citation count and open access status. Our results show that there are significant differences across databases. These differences contribute to drastic changes in rank positions of universities, which are most prevalent for non-English-speaking universities and those outside the top positions in international university rankings. Overall, MSA has greater coverage than Scopus and WoS, but with less complete affiliation metadata. We suggest that robust evaluation measures need to consider the effect of choice of data sources and recommend an approach where data from multiple sources is integrated to provide a more robust data set.

1. INTRODUCTION

Bibliometric statistics are commonly used by university leadership, governments, funders, and related industries to quantify academic performance. This in turn may define academic promotion, tenure, funding, and other functional facets of academia. This obsession with excellence is highly correlated to various negative impacts on both academic behavior and research bias (Anderson, Ronning, et al., 2007; Fanelli, 2010; van Wessel, 2016; Moore, Neylon, et al., 2017). Furthermore, these metrics (such as citation counts and impact factors) are often derived from one of the large bibliographic sources, such as Web of Science (WoS), Scopus or Google Scholar (GS). Given the potential differences between their coverages of the scholarly literature, quantitative evaluations of research based on a single database present a risky basis on which to make policy decisions.

In a related manner, these bibliographic sources and metrics are also used in various university rankings. For example, Scopus is utilized by QS University Rankings and THE World University Rankings for citation counts, while Academic Ranking of World Universities makes use of WoS for a similar purpose[1]. These rankings, and others, have been driving systematic transformations to higher education, including increased focus on student satisfaction and changes in consumer behavior. A focus on performance according to the narrow set of measures reflected in university rankings comes with a number of side effects, such as institutional homogenization, distorting disciplinary balance, and altering institutional focus (Shin & Toutkoushian, 2011; Hazelkorn, 2007). As a result of heavy criticism by the scientific community, university rankings (together with impact factors) have recently been boycotted by some academic stakeholders (Stergiou & Lessenich, 2014). This also includes domestic rankings[2]. Nevertheless, they are still widely marketed and used, without necessarily being carefully comprehended by decision makers (e.g., policymakers and students).

Bibliographic data sources evidently make a significant impact on the academic landscape. This makes the selection and use of such databases essential to various stakeholders. As such, a number of important research questions arise:

1. Are there differences across bibliographic databases?
2. If there are differences, can we characterize them?
3. Do these differences matter? How do they matter?
4. And to whom do these differences matter?

Answers to these questions may shed light on better and more robust ways to understand scholarly outputs. For all of these questions our concern is how these different analytical instruments differ in the completeness, comparability, and precision of information they provide at the institutional level. Our focus is not on reconstructing a "true" view of scholarly outputs but on a comparison of this set of tools.

1.1. Literature Review

Citation indexing of academic publications began in the 1960s, with the introduction of the Science Citation Index (SCI) by Eugene Garfield. This was followed by the annual release, starting from 1975, of impact factors through journal citation reports. This was initially developed to select additional journals for inclusion in the SCI. At that stage, much of the citation extraction was done manually (e.g., using punched cards as input to primitive computers) and the results were restricted to a niche selection of articles and journals. However, with the explosion of the Internet in the 1990s, citation indexing became automated and led to the creation of CiteSeer (Giles, Bollacker, & Lawrence, 1998), the first automatic public citation indexing system.

The rapid up-scaling of citation records created opportunities for new research explorations and bibliographic services. The former are often driven by citation analysis in the fields of bibliometrics and scientometrics, where quantitative evaluations of the academic literature play major roles. The latter is evidenced by the rise of large bibliographic and citation databases. Some of the most popular databases include WoS, Scopus, GS, and, more recently, Microsoft Academic (MSA).

WoS was the only systematic source for citation counts until 2004, when Scopus and GS were introduced. One of the earliest comparisons of these three sources was done by Jacsó (2005). The article reported on search results for citations to an article, citations to a journal,

[1] See https://www.topuniversities.com/qs-world-university-rankings/methodology, https://www.timeshighereducation.com/world-university-rankings/methodology-world-university-rankings-2018, and http://www.shanghairanking.com/ARWU-Methodology-2017.html.

[2] See, for example, http://www.bbk.ac.uk/news/league-tables.

and citations to top 30 most cited papers in a particular journal. At that time, WoS had the highest number of records simply because of its longer time span, Scopus had the widest coverage for more recent years, and GS had the lowest number of records, with limited search functions and incoherent metadata records.

Other early studies showed that Scopus offered 20% more coverage (than WoS) in citations, while GS (although with good coverage) had issues with having a shorter reference list, less frequent updates, and duplicate references in its results (Falagas, Pitsouni, et al., 2008). A number of studies have shown that the average citation counts across disciplines varied by source (Bakkalbasi et al., 2006; Kulkarni, et al., 2009; Yang & Meho, 2006). It was also shown that, for a small list of researchers, the *h*-index calculated from these three sources gave very different results (Bar-Ilan, 2008). The latest large-scale comparison showed that GS had significantly more coverage of citations than WoS and Scopus, although the rank correlations were high (Martín-Martín, Orduna-Malea, et al., 2018). Interestingly, Archambault, Campbell, et al. (2009) also showed that the rankings of countries by number of papers and citations were highly correlated between results extracted separately from WoS and Scopus.

Mongeon & Paul-Hus (2016) found that the journal coverages of both WoS and Scopus were biased toward Natural Sciences, Engineering and Biomedical Research. More importantly, their overall coverages differed significantly. Similar findings were obtained by Harzing & Alakangas (2016) when GS was added to the comparison, although for a much smaller sample of objects. Franceschini et al. (2016) also studied database errors in both Scopus and WoS, and found that the distributions of errors were very different between these two sources.

MSA was relaunched (in beta version) in 2016 as the newly improved incarnation of the outdated Microsoft Academic Services. MSA obtains bibliographic data through web pages crawled by Bing. MSA's emergence and fast growth (at a rate of 1.3 million records per month, according to Hug & Brändle, [2017]) has spurred its use in several bibliometrics studies (De Domenico, Omodei, & Arenas, 2016; Effendy & Yap, 2017; Portenoy & West, 2017; Portenoy, Hullman, & West, 2016; Sandulescu & Chiru, 2016; Wesley-Smith, Bergstrom, & West, 2016; Vaccario, et al., 2017). At the same time, various papers have tracked changes in the MSA database and compared it to other bibliographic sources (Harzing, 2016; Harzing & Alakangas, 2017a, 2017b; Hug & Brändle, 2017; Paszcza, 2016). Its rapid development, especially in correcting some important errors, over the past two years and strength in coverage have been very encouraging. However, there remain concerns regarding the accuracy of MSA's affiliation metadata. Ranjbar-Sahraei, van Eck, and de Jong (2018) found that a considerable number of publications in MSA have missing or incorrect affiliation information for a sample of output for a single university.

Tsay, Wu, and Tseng (2017) indicated that MSA had similar coverage to GS and the Astrophysics Data System for publications of a sample of physics Nobel laureates from 2001 to 2013, with MSA having a much lower internal overlap percentage than that of GS. MSA has also recently been used to predict Article Influence scores for open access (OA) journals (Norlander, Li, & West, 2018). Hug, Ochsner, and Brändle (2017) and Thelwall (2018), using samples of publications, showed there was uniformity between citation analyses done via MSA and Scopus. Harzing & Alakangas (2017a) also showed, for individual researchers, that the citation counts by MSA were similar to or higher than Scopus and WoS, varying across disciplines.

1.2. What Is Different in This Study?

As discussed by Neylon & Wu (2009), using a singular article-level or journal-level metric as a filter for scientific literature is deeply flawed and incorporating diverse effective measurement tools is a necessary practice. In a similar vein, using a single bibliographic data source for evaluating specific aspects of academic work can be misleading. Given the immense social and academic impacts of the results of such evaluations, and the unlikeliness of them (as either part of research quantification or rankings) being completely discarded any time soon, one ought to be cautious in both interpreting and constructing such evaluation frameworks. With this in mind, we aim to provide a deep exploration in comparing the coverage of research objects with DOIs (digital object identifiers) in WoS, Scopus, and MSA[3], in terms of both volume and various bibliographic variables, at the institutional level. In particular, a sample list of 15 universities is selected (ranging in geography, prestige, and size) and data affiliated with each university are drawn from all three sources (from 2000 to 2018). Less detailed data are also collected for another 140 universities to be used as a supplementary set where applicable. An automated process is used to compare the coverage of the sources and the discrepancies in the publication year are recorded. In addition, manual online searches are used to validate affiliation correctness and plausibility for samples of DOIs. The focus on DOIs also provides broader opportunities for cross-validation of bibliographic variables, such as OA status and document types from Unpaywall[4], and citations data from OpenCitations[5]. These will assist in further understanding of the differences between data sources and the kind of biases that they may lead to.

Previous studies that compared WoS, Scopus, and MSA were limited to publications linked to an individual researcher, a small group of researchers, or one university. These comparisons were also mostly drawn in relation to citation counts. This article extends the literature by expanding the study set to include several universities and drawing institutional comparisons across a larger selection of characteristics and measures. We further sought to distinguish between the effects of completeness of research outputs in different data sources from the effects of that coverage on any specific performance metric (such as citation counts). Most previous studies have focused on the question of applying a specific data source to the specific problem of creating a citation-based ranking. An important difference for us was to ask the more general question of how bibliographic sources might affect a range of different performance evaluations, including, but not limited to, citations as the performance metric.

The study therefore includes analyses of potential effects in the exclusive selection of one source for evaluating a set of bibliographic metrics (i.e., potential effects on the ranking of universities). We use secondary data sources (Unpaywall and OpenCitations) to construct metrics for OA and citations. This gives standardized contrasting sets of records for comparisons across bibliographic sources. In turn, this allows us to disentangle the questions of research output completeness, and the separate effects this may have on quantitative measures. The results lead up to the main message that it is essential to integrate diverse data sources in any institutional evaluation framework.

The remainder of this article is structured as follows: Section 2 gives an overview of some global characteristics across the various bibliographic databases. Section 3 provides detailed descriptions of our data collection and manual cross-validation processes. All analyses and

[3] We have selected WoS, Scopus, and MSA for our analysis because they provide structured metadata handling and comprehensive API search functions. GS is not considered due to difficulties in metadata handling and lack of API support, but it may be of interest for examination in future work (especially given the apparent large scale of coverage).

[4] https://unpaywall.org/.

[5] https://opencitations.net/.

results are presented in section 4. Sections 5 and 6 contain discussions on limitations and conclusions, respectively.

2. GLOBAL COMPARISON OF FEATURES AND CHARACTERISTICS ACROSS WOS, SCOPUS, AND MSA

WoS and Scopus are both online subscription-based academic indexing services. WoS was originally produced by the Institute for Scientific Information (ISI), but was later acquired by Thomson Reuters and then Clarivate Analytics (formerly a part of Thomson Reuters). It contains a list of several databases, where the level of access to each depends on the selection of subscription models. The search functionalities can also vary according to which databases are selected (for example, the "Organization-Enhanced" search option is not available when all WoS databases are included). On the other hand, Elsevier's Scopus database seems to offer one unified database of all document types (the only exception is data on patents, which appear as a separate list in search results). A manual online search reveals a wider variety of document types in WoS. For example, it contains items listed as "poetry," which does not seem to fit into any of the types in Scopus.

MSA is open to the public through the Academic Knowledge API, though both a rate limit and a monthly usage cap apply to this free version[6]. The subscription version is documented as relatively cheap at \$0.32 per 1,000 transactions[7]. Its semantic search functionality and ability to cater for natural language queries are among the main differences from the other two bibliographic sources. Its coverage in patents has greatly increased through the recent inclusion of Lens.org metadata[8]. As a preliminary examination, we take a look at some global characteristics and features across the three sources. Table 1 provides an overview of coverage and comparative strengths in each source. WoS has several databases from which it extracts data. The most commonly used collection of databases is WoS Core, which allows for more functionality. On the other hand, WoS All Databases includes all databases listed by WoS (with increased coverage for Social Sciences and local languages, for example), but due to varying levels of availability of information its functionalities are limited (e.g., fewer search query options). Scopus does not seem to index Arts & Humanities, while MSA appears to have significantly more coverage in Social Sciences and Arts & Humanities than WoS Core and Scopus. With higher coverage for journals and conferences, MSA tracks a significantly larger set of records. It is also interesting to note that MSA had approximately 127 million documents only a couple of years ago (Herrmannova & Knoth, 2016).

The annual total numbers[9] of objects for the various sources from 1970 to 2017 are displayed in Figure 1. In comparison to Jacsó (2005), and other studies mentioned earlier, there seem to be significant increases in both Scopus and WoS in terms of both growth over time and backfilling. However, both sources still have significantly less total counts than that of MSA. The figure also shows a high degree of correlation between Scopus, WoS Core, and WoS All Databases. However, this figure does not provide any information on internal or external overlaps across the sources (which we shall explore).

To get a better overview of research disciplines covered by each source, the percentage spread of objects across disciplines, for each source, is displayed in Figure 2. Evidently, all

[6] See https://dev.labs.cognitive.microsoft.com/products/5636d970e597ed0690ac1b3f.

[7] See https://azure.microsoft.com/en-au/pricing/details/cognitive-services/academic-knowledge-api/.

[8] See https://www.microsoft.com/en-us/research/project/academic/articles/sharpening-insights-into-the-innovation-landscape-with-a-new-approach-to-patents/.

[9] Publication year defined as per source.

Table 1. Coverage and features of WoS[10], Scopus, and MSA

	WoS Core	WoS All	Scopus	MSA
Subject focus	Sciences and Technology, with some coverage for Social Sciences and Arts & Humanities	Similar to WoS Core, but with significant increases in Social Sciences	All Sciences, with some coverage for Social Sciences and Arts & Humanities	Sciences, but with significantly more coverage for Social Sciences and Arts & Humanities
Total count[11]	62,602,202	73,808,358	72,170,639	206,252,196
Time span[12]	From 1972	From 1950	From 1858	From 1800
Coverage[13]	>20,300 journals, books and conference proceedings	>34,200 journals, books, proceedings, patents and data sets	24,130 journals; 245 conference series; includes books, book chapters and patents	47,989 journals; 4,029 conference series; includes books, book chapters and patents
Updating frequency	Daily (Monday to Friday)	Ranges from daily to monthly	Daily	Weekly
Strengths	• Comprehensive search options: affiliation, DOIs, year, etc. • Organization-enhanced search • Provides detailed OA information as per Unpaywall	• Provides detailed OA information as per Unpaywall • Provides coverage of patents and data sets • Increased regional coverage (e.g., Russia, Latin America and China)	• Simple API access available through Python • Comprehensive search options: affiliation, DOI, year, etc.	• API available through languages such as R and Python • Strong coverage of all subject areas • Includes patents data from Lens.org • Semantic search
Weaknesses	• Limited coverage for Arts & Humanities	• Does not seem to provide affiliation search (and various other queries that are available for WoS Core)	• Fewer details of OA information are provided[12] • Limited coverage of Arts & Humanities	• No apparent record of OA information • Very few search options through the web service • Completeness and accuracy of metadata less studied

[10] It is important to note that our access to WoS is dependent on the WoS license of our institution. This determines which indices are included and the time range for each. See Supplementary Material 2 for further detail.

[11] As per website search or report on August 7, 2018. Numbers reported are not necessarily the same as the total number of user-accessible records. For estimates of user-accessible records, see Gusenbauer (2019).

[12] As permitted through the advanced search functions in WoS and Scopus on August 7, 2018.

[13] While this article was being prepared, Elsevier announced their agreement to use Unpaywall data (see https://www.elsevier.com/connect/elsevier-impactstory-agreement-will-make-open-access-articles-easier-to-find-on-scopus) and later implemented it (https://blog.scopus.com/posts/scopus-makes-millions-of-open-access-articles-easily-discoverable).

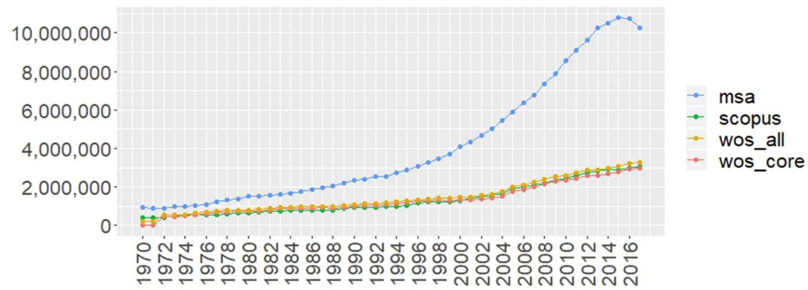

Figure 1. Annual total item counts for Scopus, MSA, WoS Core, and WoS All from 1970 to 2017.[14]

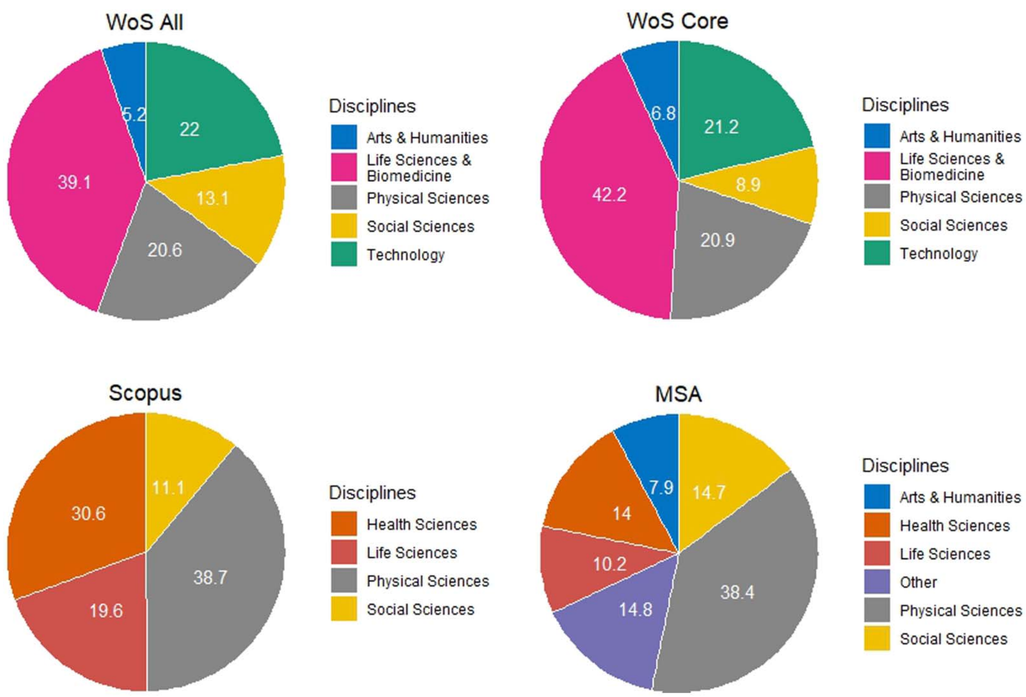

Figure 2. Distributions of objects in WoS All[15], WoS Core[16], Scopus[17], and MSA[18] among disciplines.

[14] Data as of August 15, 2019.

[15] This is for all databases in WoS. Counts were obtained by querying for all research areas under each of the five broad categories (as defined by WoS) using the Advanced Search function on WoS, as at August 3, 2018.

[16] This is for the core databases in WoS. Counts were obtained by querying for all research areas under each of the five broad categories (as defined by WoS) using the Advanced Search function on WoS, as at August 8, 2018.

[17] Obtained by querying for each broad subject area (as defined by Scopus) through Scopus' Advanced Search option, as at August 3, 2018.

[18] MSA did not seem to have broadly defined disciplines. Counts for the 19 top-level fields of study were obtained from the Topics Analytics page (on August 3, 2018). Then we sorted their detailed disciplines into broader ones (roughly following those in WoS) as follows: Health Sciences = Medicine; Physical Sciences = Chemistry, Engineering, Computer Science, Physics, Materials Science, Mathematics, Geology; Life Sciences = Biology, Environmental Science; Social Sciences = Psychology, Geography, Sociology, Political Science, Business, Economics; Arts & Humanities = History, Art, Philosophy.

sources are dominated by the sciences, as commonly noted in the literature. However, MSA does seem to have relatively higher proportions for both Social Sciences and Arts & Humanities.

3. METHODOLOGY AND DATA

3.1. Methodology

To perform a more detailed comparison of sources, we gathered outputs for a selected set of 15 universities, ranging in geography, prestige, and size, from each bibliographic source: WoS Core (referred to as just WoS from here on), Scopus, and MSA. This is done through the use of APIs for each data source. We extract records for the years from 2000 to 2018 via affiliation IDs (in the case of Scopus and MSA) and organization-enhanced search terms (for WoS)[19]. The results form three sets of data (one from each source) for each university. Subsequently, DOIs of objects (for those that do have them) are extracted from each set. A further 140 universities are also included as a supplementary set to be used where necessary. Our strategy is to explore various bibliographic characteristics related to these DOIs at the overall level (for all years and all institutions) and then contrast that with the corresponding results for individual universities focusing on a single year (i.e., 2016). Where applicable, the analysis for 2016 is also extended to the full set of 155 universities. We are mainly interested in the following characteristics:

1. Distributions (e.g., Venn diagrams) of DOIs across sources,
2. Discrepancies in publication year recorded by each source,
3. Document types across various parts of the Venn diagrams of DOIs,
4. Citation counts (as per OpenCitations) calculated across sources,
5. OA levels (as per Unpaywall) calculated across sources, and
6. Plausibility of assigned affiliation for DOIs exclusively indexed by a single source.

Analyses of characteristics 1 to 5 are mostly automated, with data collected into a data management systems implemented on the Google Cloud Platform. Data are gathered via the APIs of WoS[20], Scopus, and MSA. Specifically, each data source is queried for a specified time period for articles with specified affiliation. Our collection process is therefore dependent on both the affiliation and publication year metadata for each source. For example, a DOI may be indexed in both source A and source B, where it is assigned to university X by source A but not by source B. This DOI will not be retrieved from source B as an output from university X, even though it is indexed in source B. Similarly a significant number of DOIs are assigned to different publication years by different sources. Hence, the resulting Venn diagrams do not show the overall coverage of all research output for a given institution, but rather the discoverability via DOI, publication year, and affiliation.

Conceptually, we use each source to prepare the output data for 18 editions of an annually constructed evaluation or ranking. To examine the overlap between sources for our initial search criteria (i.e., affiliation and publication year), Venn diagrams for characteristic 1 are constructed by matching DOIs and publication years recorded by each source.

To explore the potential reasons for discrepancies across sources, we also examine characteristics 2 and 3. Metadata of each DOI from each source are compared to determine the level of agreement in publication years across sources. The “genre” field from Unpaywall is used to determine the document type of each DOI (e.g., journal articles, book chapters, conference proceedings).

[19] See Supplementary Material 1.

[20] See Supplementary Material 2 for a list of WoS databases accessed in this study.

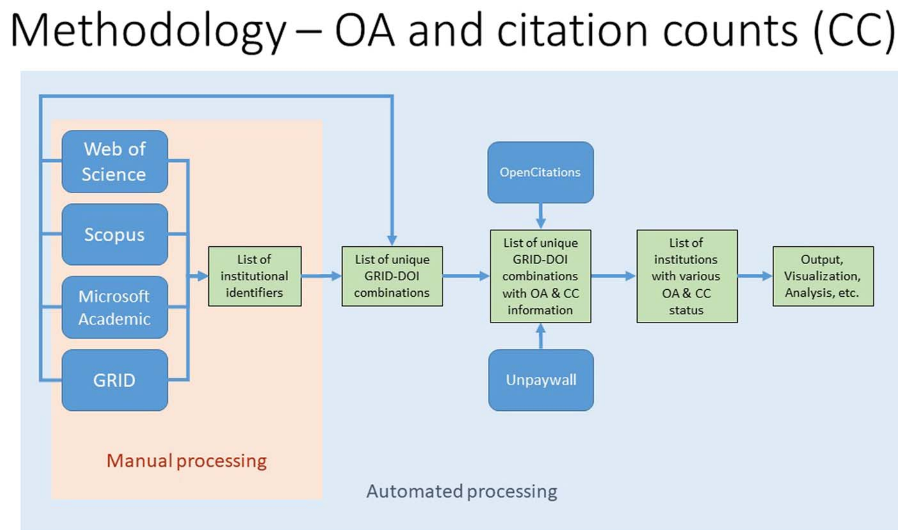

Figure 3. A summary of the data collection process.

Unpaywall data and OpenCitations' COCI data are used to determine OA status and citation counts associated with each DOI. For this article, we only require the general OA status and not the type of OA (e.g., gold OA, green OA). Hence, we only use the "is_oa" field in the Unpaywall metadata to determine the OA status of DOIs in our data. COCI records citation links between Crossref DOIs. By querying and merging all links to a DOI, it allows us to determine the number of citations this DOI receives. The use of COCI data allows us to separate the effects of metadata coverage (i.e., publication and affiliation metadata) from the effects of inclusion of the sources of citations in a specific database. While COCI includes only those citations that are made openly available through Crossref (and therefore excludes citations from a number of large publishers, including Elsevier and the American Chemical Society at the time of writing) it nonetheless provides us with a consistent source of data that can be applied to the evaluation of all DOIs. We focus here on the differences in coverage among bibliographic data sources, and how this affects the results of evaluations based on other data sources. This is different from seeing how inclusion in a citation database affects the result of a citation-based ranking. That is, our goal is to compare the use of these data sources as instruments for discovering outputs that might then be evaluated by a range of external measures, rather than as instruments for constructing self-contained citation-based evaluations. For comparison, we also provide an evaluation of the differences in rankings constructed using citation data from each source in Supplementary Material 9.

We gather this information for a set of DOIs of interest (e.g., DOIs from WoS affiliated to one university) and obtain total citation counts for this set. This total can then be divided by the number of (Crossref) DOIs affiliated to this university to produce an average citation count[21]. The overall data collection process is summarized in Figure 3[22]. The codes and data to produce the results and figures in this article can be accessed at Huang, Neylon, et al. (2019).

[21] This implies that a Crossref/Unpaywall DOI that does not have any inward citation links in COCI is assumed to have zero citation for this study.

[22] We use affiliation IDs from the Global Research Identifier Database (GRID: https://www.grid.ac/) as the standardized identifier for each institution. These are mapped to IDs and search terms in WoS, Scopus, and MSA as shown in Supplementary Material 1. Much of the mapping of institutional identifiers is manually processed at this stage.

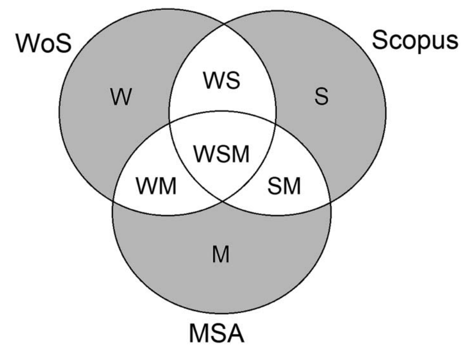

Figure 4. Nonoverlapping sections (in grey) of the spread of DOIs from three sources for an institution in a particular year.

A manual process is followed for checking characteristic 6. The procedure for the manual validation is focused on the nonoverlapping parts of the three sources (i.e., shaded sections in Figure 4). The overlapping parts indicate agreement by at least two sources over both affiliation and publication year records (when filtered down to a particular year). Given the different ways in which the sources gather data, the reliability of information for these parts is much more convincing. In contrast, the nonoverlapping sections are not validated by other sources. This leads to the need for the manual validation process.

The publication year can be a reason for the discrepancy of coverage due to inconsistencies in how date is recorded. For example, in the case of a journal article, a source may choose to record the date of the journal issue, the publication date for the article, or the date on which the article first appeared online. Hence, our first step is to check whether DOIs from the nonoverlapping sections are indeed in another source but fall in a different year. After removing these DOIs, which were identified via comparison to adjacent years, we sampled the remaining DOIs from each nonoverlapping section for manual validation (Figure 4). This is processed for DOIs from 2016.

The process that leads to the manual validation is summarized in the flowchart given in Figure 5. Once DOIs are sampled from each nonoverlapping section, they are compared against the other two sources (via DOI and title searches on each source's webpage) and also the original document (online versions)[23].

Assume we have three sources, A, B, and C, and the current set of DOIs are from source A. The following questions are asked as part of the manual checking process (with a likewise procedure used for DOIs from the other two sources):

1. Is this DOI found in the metadata record in source B?
2. Is the title associated with this DOI found in source B?
3. Is the exact affiliation phrase found in the metadata record in source B?
4. If not, is the affiliation plausible?
5. Is this DOI found in the metadata record in source C?
6. Is the title associated with this DOI found in source C?
7. Is the exact affiliation phrase found in the metadata record in source C?
8. If not, is the affiliation plausible?
9. Is the DOI correctly recorded in source A (as per original document or doi.org)?
10. Is the exact affiliation phrase found on the original document?
11. If not, is the affiliation plausible?

[23] This cross-validation process was carried out manually by a data wrangler, on a part-time basis over a few months, for which online data was accessed from December 18, 2018 to May 20, 2019.

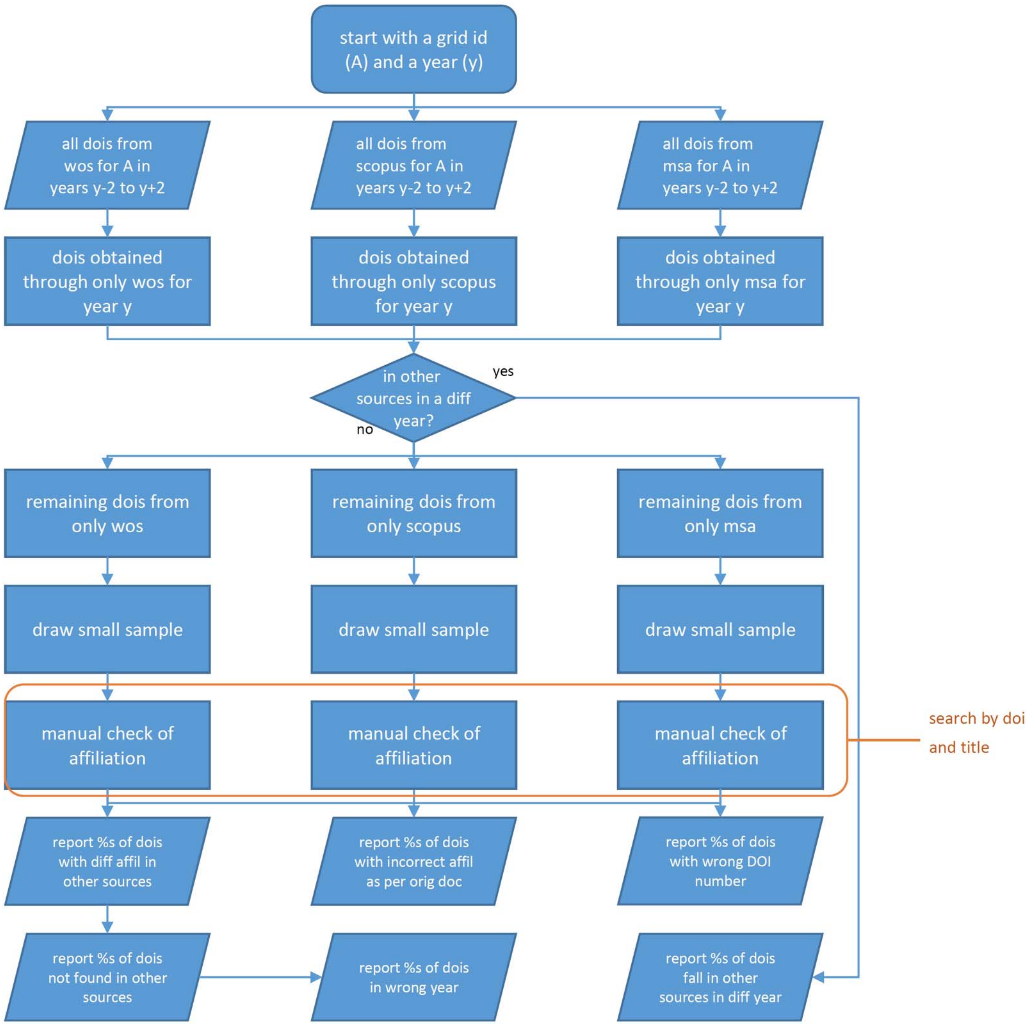

Figure 5. Flowchart of the process leading to manual validation.

The numbers of DOIs to be sampled for each institution are 30, 30, and 40 from (exclusively) WoS, Scopus, and MSA, respectively, after removal of DOIs that are found in another source for a different year.

3.2. Data

Table 2 presents the total number of unique DOI records we have obtained from each source, the combined number of unique DOIs, and how many of these DOIs are recorded in Unpaywall, for each institution for 2016. The coverage of DOIs by Unpaywall is very high, as expected. The only slight exception is DUT, where a significantly higher portion of Scopus DOIs were not recorded by Unpaywall. A quick exploration[24] finds most of these DOIs to be registered with China National Knowledge Infrastructure (CNKI) or the Institute of Scientific and Technical Information of China (ISTIC), whereas Unpaywall currently only indexes Crossref DOIs[25].

[24] Using the Crossref API for agency information.
[25] See https://unpaywall.org/user-guides/research.

Table 2. The spread of DOIs (for 2016) across various sources for 15 different institutions and percentages of them recorded in Unpaywall

	WoS		Scopus		MSA		Combined	
Institution[26]	Total count	Unpaywall[27]	Total count	Unpaywall	Total count	Unpaywall	Total count	Unpaywall
Cairo	2,629	98.4%	2,454	98.7%	2,761	98.7%	3,793	97.4%
Curtin	3,168	99.2%	2,963	98.8%	3,150	98.8%	4,070	98.2%
DUT	3,552	97.9%	3,789	82.8%	3,582	99.5%	5,091	86.1%
IISC	2,002	98.6%	1,964	98.9%	2,464	97.8%	3,156	97.8%
ITB	585	99.1%	1,040	99.7%	1,027	96.8%	1,744	97.8%
LU	1,419	97.8%	1,357	98.6%	1,564	99.2%	2,014	97.5%
MIT	8,053	99.3%	6,702	99.1%	7,457	97.6%	10,889	98.6%
MSU	5,480	99.2%	5,107	99.5%	5,719	99.3%	7,362	98.8%
UNAM	4,754	98.6%	4,258	98.8%	5,401	96.8%	7,056	96.5%
UCL	13,615	99.0%	11,255	99.1%	9,924	98.7%	17,230	98.4%
UCT	3,079	98.8%	2,852	98.8%	3,189	98.3%	4,206	97.9%
Giessen	1,871	98.7%	1,638	99.4%	1,545	99.2%	2,354	98.2%
USP	11,451	98.3%	10,923	99.4%	13,664	96.8%	17,732	96.5%
Tokyo	9,640	99.0%	8,810	99.1%	9,789	97.9%	12,848	97.9%
WSU	2,717	98.7%	2,041	99.2%	2,794	99.1%	3,569	98.2%
Overall[28]	71,709	98.8%	65,017	98.2%	72,386	98.5%	100,456	97.2%

4. ANALYSIS AND DISCUSSION

In this section, we proceed with the detailed bibliographic comparisons across sources. We will start with exploring the coverage of DOIs by each source. This is followed by examining the amount of agreement, or disagreement, of publication year recorded by each bibliographic data source. The document types, citation counts, and OA percentages, as per source, are the subsequent analyses. Lastly, a manual cross-validation procedure is employed for samples extracted from nonoverlapping sections of the Venn diagrams for each institution in our sample of 15 institutions.

4.1. Coverage and Distribution of DOIs

Here we take an exploration of the spread of the DOIs across the sources. Figure 6 shows the Venn diagrams of DOI counts for our initial set of 15 universities combined from 2000 to 2018 and for just 2016[29], respectively. Evidently, the central regions (overlap of all three sources) have

[26] Cairo University, Curtin University, Dalian University of Technology (DUT), Indian Institute of Science Bangalore (IISC), Institut Teknologi Bandung (ITB), Loughborough University (LU), Massachusetts Institute of Technology (MIT), Moscow State University (MSU), National Autonomous University of Mexico (UNAM), University College London (UCL), University of Cape Town (UCT), University of Giessen, University of Sao Paulo (USP), University of Tokyo, Wayne State University (WSU).

[27] The number of DOIs that are indexed by Unpaywall.

[28] The set of unique DOIs for all 15 institutions combined.

[29] Dates as per source's metadata.

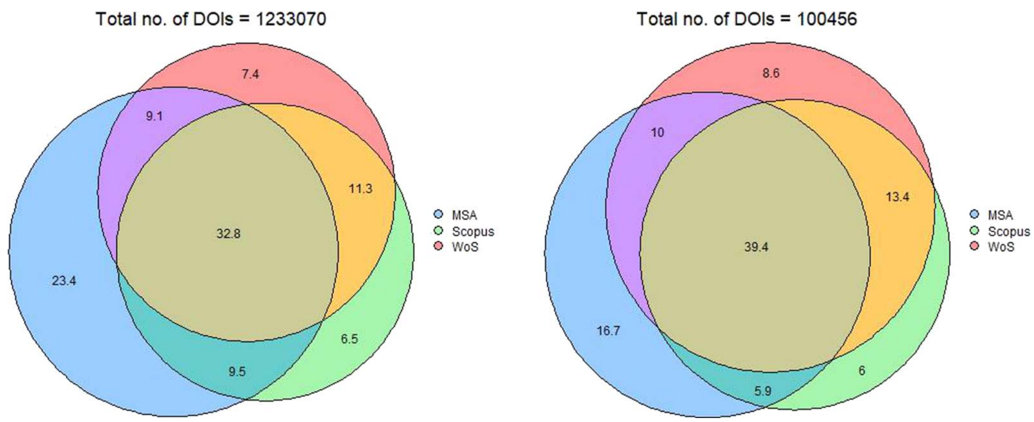

Figure 6. Percentage Venn diagrams of DOIs from all 15 institutions for years 2000–2018 (left) and for only 2016 (right).[30]

the highest count in each Venn diagram. These are DOIs that have been indexed by all three sources and, given the intended global coverage of major publication venues by these sources, the relatively higher counts here are not at all surprising. However, there are also significant portions of DOIs exclusively accessed via a single source in both Venn diagrams. This gives rise to the potential biases in any bibliometric measure to be calculated from a single source.

This pattern of difference in coverage is mirrored at the institutional level, albeit the level of discrepancies varies across institutions. Supplementary Material 3 contains two Venn diagrams for each institution, both for 2016. In each case, the Venn diagram on the left records all DOIs as per bibliographic source and the one on the right is a subset of these DOIs that are also indexed in the Unpaywall database. It is noted that the two Venn diagrams for each institution are quite similar due to the high coverage of these DOIs by Unpaywall. The only exception is the Scopus coverage of DOIs for DUT, for which the DOIs exclusively indexed by Scopus significantly decreased when moving from the left Venn diagram to the one on the right. This is consistent with what we observed earlier, with many of these DOIs provided by agencies other than Crossref. The overall pattern is that there appear to be significant portions of DOIs only indexed by a single source. Hence, pulling together these sources can greatly enhance coverage. Interestingly, for most institutions, MSA has the most number of exclusively indexed DOIs (and appears to be more extreme for non-English-speaking and non-European universities), the only exception being UCL. ITB also represent an extreme case where the proportion of DOIs indexed by all three sources is much lower relative to other universities.

To have a better overview of how coverage of these three sources varies across institutions, we perform several analyses as follows. First, we identify each institution with the seven different counts (instead of percentages) as per its own Venn diagram of all DOIs (Venn diagrams on the left in Supplementary Material 3). We also include another 140[31] universities for comparison. We view each (GRID ID, DOI) pair as a distinct object. Hence, we obtain a 155 × 7 contingency table. Each column of this table represents the number of DOIs falling in the respective section of the Venn diagram; for example, column 1 is the number of DOIs in section

[30] Readers are reminded that these Venn diagrams do not attempt to show the true coverage of all research output affiliated to these universities by each source. Rather, they represent the discoverability of research output with DOIs that are linked to the affiliations and time frames of interest by each source.

[31] Originally, there were 150 additional universities, but 10 were removed due to noncoverage or identification issues (e.g., multiple Scopus affiliation IDs). See Supplementary Material 1 for the list of GRID IDs of the additional 140 universities.

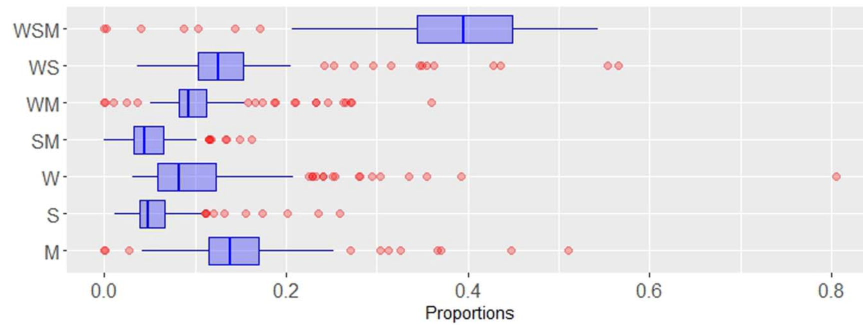

Figure 7. Boxplots of proportions of DOIs that fall in each section of the Venn diagram across 155 universities for 2016.

WSM of the Venn diagram (refer to Figure 4). We can also convert these counts to proportions by dividing them by the total number of DOIs for each institution. Figure 7 shows the distribution of these proportions for each section of the Venn diagram across all 155 universities. The higher proportion in the central region (section WSM) of the Venn diagram is again observed. The general pattern that emerged is that, for all sections of the Venn diagram, there appears to be a concentrated central location with many extreme cases (excess kurtosis of 2.29, 9.72, 5.96, 1.82, 22.24, 11.49, and 6.88, from sections WSM, WS, WM, SM, W, S, and M, respectively) and substantial skewness. We can also concatenate the respective sections to get the proportion of DOIs covered by each bibliographic source. The spreads of these proportions are summarized in Figure 8 as histograms.

Again, the pattern of high central peak, skewness, and long tail is observed. The peakedness and long tails are confirmed by the excess kurtosis of 4.29, 3.34, and 8.60 for WoS, Scopus, and MSA respectively. The skewness to the left with number of extreme cases highlights the low degree of coverage for some universities. Meanwhile, a correlation analysis of the proportions for the three sources is quite intriguing (see Table 3). Both Spearman's rank correlation and Pearson's correlation matrices are presented here. There appears to be a negative correlation between coverage by WoS and coverage by MSA: When there is a high proportion of

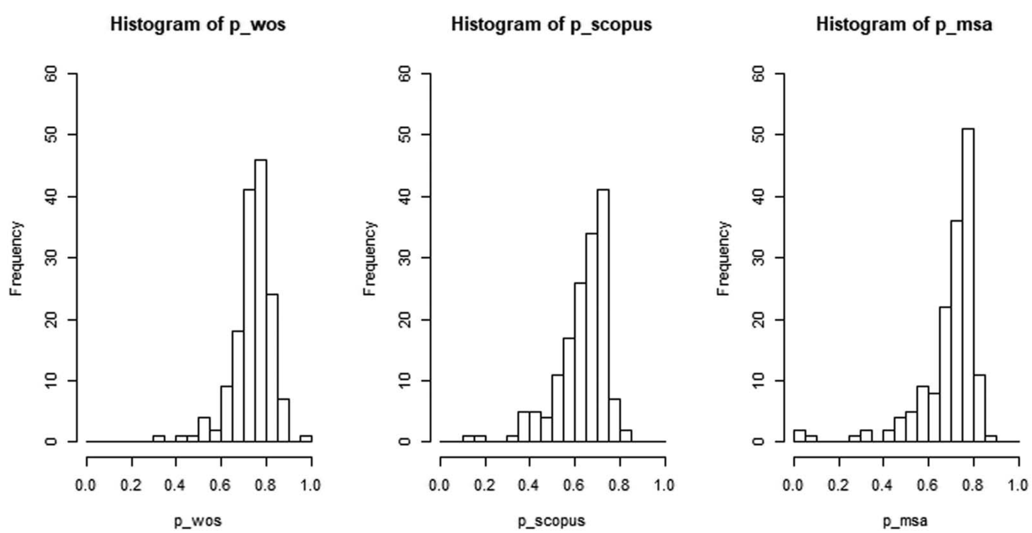

Figure 8. Histograms of proportions of DOIs in WoS, Scopus, and MSA for 2016 (across 155 universities).

Table 3. Spearman's rank correlation and Pearson's correlation matrices of proportions of DOIs covered by each bibliographic source

	Spearman			Pearson		
Sources	**p_wos**	**p_scopus**	**p_msa**	**p_wos**	**p_scopus**	**p_msa**
p_wos	1	0.07	−0.50	1	0.08	−0.39
p_scopus	0.07	1	0.10	0.08	1	0.06
p_msa	−0.50	0.10	1	−0.39	0.06	1

Table 4. *p*-values for tests of homogeneity across institutions in terms of the distribution of DOIs

Sample	Chi square[32]	Chi square MC[33]	G test
15 universities	<0.0001	0.0002	<0.0001
155 universities	<0.0001	0.0002	<0.0001

coverage by WoS, the coverage by MSA is relatively low. There is also a low correlation between WoS and Scopus. While much of this may be attributed to the different methodological structure and focus across WoS, Scopus, and MSA, the degree of nonalignments is still quite a surprise[34].

We further performed tests of homogeneity across institutions to check whether the spread of DOIs across individual Venn diagrams come from the same probability distribution. The results of these tests are provided in Table 4. It is evident that the chance of rejecting homogeneity is very high. Bootstrapped samples from sample sizes 10 to 155, in increments of 5, all gave similar results as well.

It is also expected that these Venn diagrams are not symmetrical (in the sense of equal proportional coverage across each source), which is observable from the Venn diagrams of our initial sample of 15 universities in Supplementary Material 3. However, to obtain further insight into the symmetry of a large number of Venn diagrams (i.e., all 155 universities), we introduce three related measures. Let p_i be the proportion of DOIs that fall in part i of a Venn diagram and define the following three measures:

$$d_1 = |p_{WS} - p_{SM}| + |p_{SM-} - p_{WM}| + |p_{WM} - p_{WS}| + |p_W - p_S| + |p_S - p_M| + |p_M - p_W|$$

$$d_2 = |p_{WS} - p_{SM}| + |p_{SM} - p_{WM}| + |p_{WM} - p_{WS}|$$

$$d_3 = |p_W - p_S| + |p_S - p_M| + |p_M - p_W|$$

where d_1 is the sum of absolute differences across the whole Venn diagram, d_2 is the sum of inner differences, and d_3 is the sum of differences across the outer regions of the Venn diagram. We calculate values for these three measures for each university's Venn diagram and

[32] None of the cells in these contingency tables has an expected count less than 10.

[33] Using the sampling procedure for that of Fisher's exact test with 5,000 replicates. See https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/chisq.test

[34] To see whether these correlations are driven by the size of total output, we have also constructed pairwise scatterplots between the three proportions, with the points color-coded by total output numbers. The random spread of the colors suggested the correlations are not strongly influenced by size. See Supplementary Material 4.

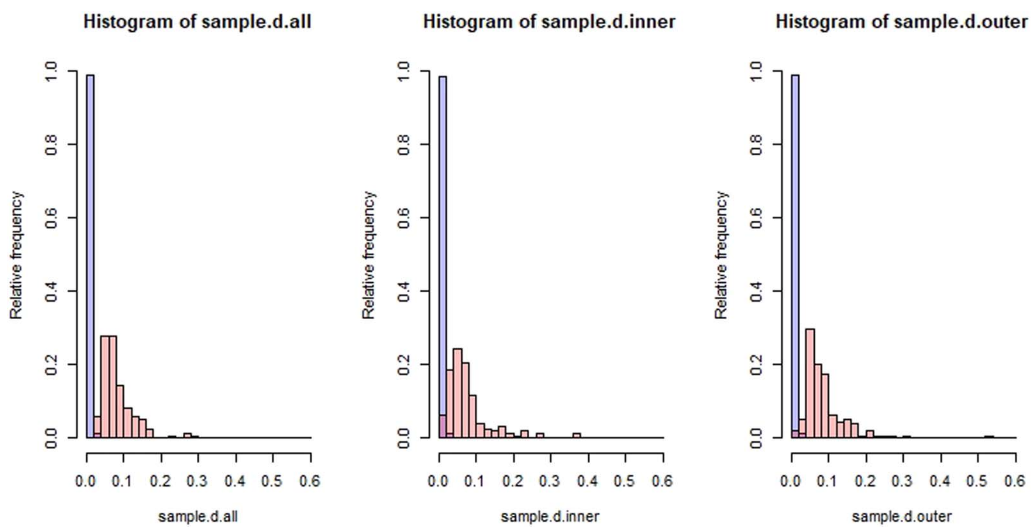

Figure 9. Histograms of d_1, d_2, and d_3 (left to right, respectively) for our data of 155 universities (in red) and for randomly generated symmetrical Venn diagrams (in purple).

compare their distributions to those produced by randomly generated Venn diagrams. Firstly, they are compared to randomly generated symmetrical Venn diagrams[35]. The resulting distributions are presented in Figure 9. It is quite obvious that the results from our data do not correspond to those generated from symmetrical Venn diagrams. As further contrasts, we also compare these measures against Venn diagrams generated from various other distributions (see Supplementary Material 5). As expected, our data are better represented by distributions other than those produced by symmetrical Venn diagrams. Furthermore, there appear to be some differences in distributions across d_1, d_2, and d_3, which we do not further examine here and leave for future exploration.

Now that we have confirmed the differences in DOI distributions across institutions and negative to low correlations between the nonsymmetrical coverages by the three bibliographic sources, a follow-up question may be whether there are groupings among these universities. We proceed with a hierarchical cluster analysis for both the sample of 15 universities and for all 155 universities, using dissimilarities between the proportions of the Venn diagrams as clustering criteria[36]. At the same time, we also color-code the universities by their regions and rank positions on the 2019 THE World University Rankings. Some of these are presented in Supplementary Material 6. While no striking patterns emerge, there do appear to be some interesting groupings. For example, there seems to be a block of European and American universities toward the left of the dendrogram colored by region. Perhaps unsurprisingly, around the same area for the dendrogram colored by THE ranking, there is also a rough cluster of the most highly ranked universities.

The contrasts may be more apparent for the smaller sample of 15 universities. An example of this is presented in Figure 10. ITB is clearly an outlier from the rest of the group, as was the case for the Venn diagrams, and the two highest ranked universities are placed quite close to

[35] p_wos = p_scopus = p_msa generated from a uniform distribution (truncated at ⅓ and 1).

[36] Hierarchical clustering is performed using `hclust` function (base `R`) with dissimilarity matrix calculated using Gower's distance in the `daisy` function (`R` package `cluster`). Graphical presentations are produced using the `R` package `dendextend`.

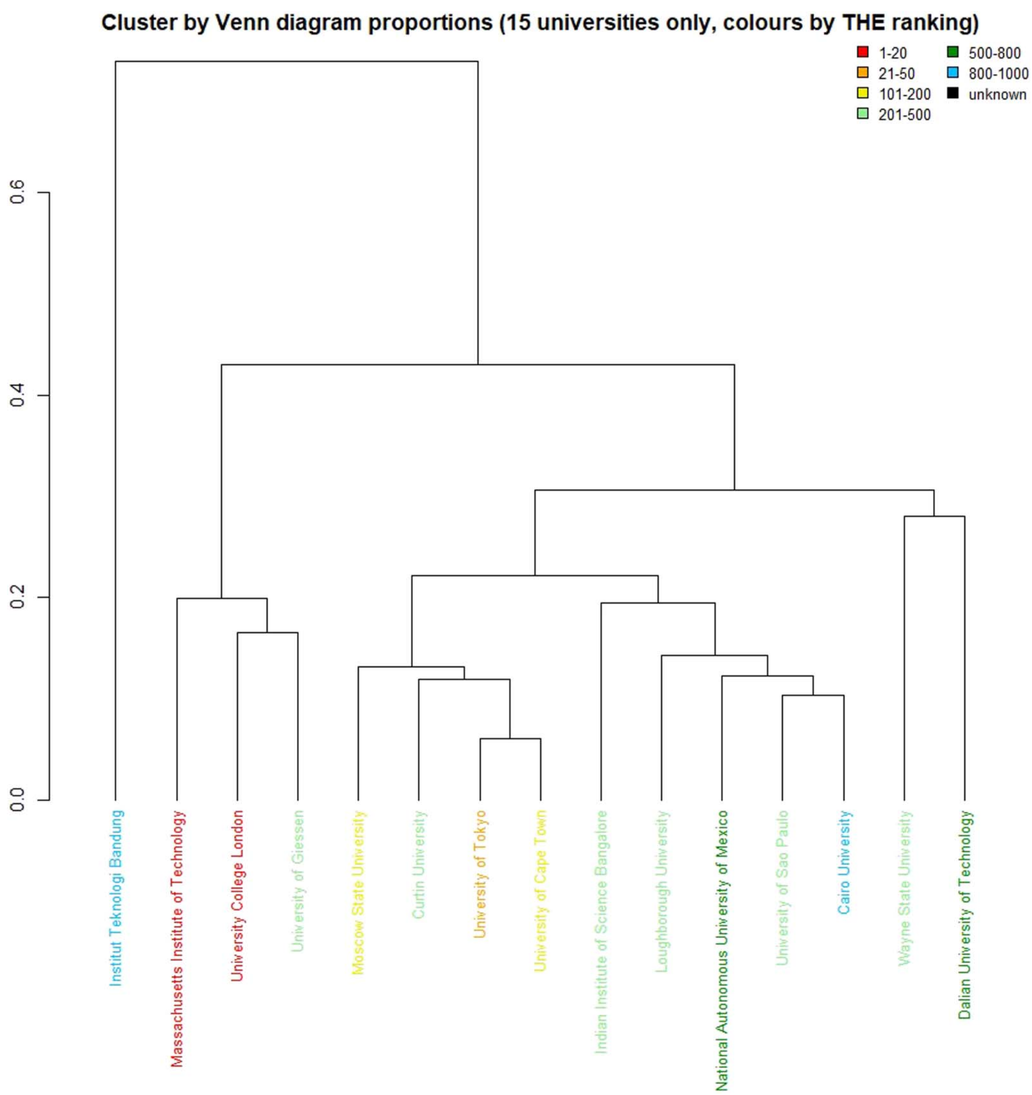

Figure 10. Dendrogram showing clustering of 15 universities by Venn diagram proportions vs rank position on 2019 THE World University Rankings.

each other. Seven of the universities ranking from 201 and above are placed on the right of the dendrogram (perhaps in two clusters). One of these also consists mainly of universities from non-English-speaking regions (Loughborough being the exception). In general, there appear to be some patterns of prestige and regional clustering (for both the sample of 15 and the sample of 155 universities).

4.2. Comparison of Publication Years

As mentioned earlier, discrepancies in publication year recorded by different bibliographic sources are possible, given that there is no universal standard to define publication year (or indeed publication date). This poses a problem when trying to combine sources to evaluate and track a bibliometric variable over time. A DOI can be double-counted (i.e., counted two or more times in different years via different sources). In the following, we explore the amount of agreement and disagreement on publication years by WoS, Scopus, and MSA. The overall

Table 5. Agreement on "publication year" across bibliographic sources for 15 universities combined (from 2000 to 2018)

Sources	WoS/Scopus	WoS/MSA	Scopus/MSA	All three sources
Total agreement on year	541,501	508,269	509,797	397,560
Total overlap of DOIs[37]	544,133	516,986	522,026	404,710
% of agreement	99.5%	98.3%	97.7%	98.2%

numbers are presented in Table 5, covering all DOIs for 15 institutions and years from 2000 to 2018.

In this table, the number of DOIs jointly indexed by pairs of bibliographic sources and by all three bibliographic sources are recorded. It should be noted that percentages are calculated over different sets of DOIs (i.e., different denominators). For example, the number of DOIs common to all three sources (i.e., 404,710) is less than the number of DOIs common to only Scopus and MSA (i.e., 522,026).

It is clear that the overall levels of agreement are very high. However, two follow-up questions are (1) for DOIs that are present in a different source for a different year, what is the spread of these DOIs over years? And (2) while the overall agreement of publication years is high, does that carry over to individual institutions?

To answer these questions, we now focus our attention on the year 2016 and DOIs that are exclusively indexed by a single source for that year. Figure 11 displays the spread of such DOIs from a particular source when matched against the other sources for different years. These are again DOIs from our sample of 15 institutions combined. The majority of the discrepancies are within one year (i.e., falling in 2015 and 2017), while extending this window one further year in both directions covers almost all remaining cases. We also note some differences across the sources. The number of discrepancies between WoS and Scopus is relatively small compared to those involving MSA. This may be the likely result of MSA using the date when a document first appears online as their default publication date[38].

Next we explore how these discrepancies of the publication year are distributed for individual institutions. Table 6 records, for each source, the percentages of DOIs from 2016 that appear in the other two sources but differ by a year and two years, respectively. For WoS, the percentage of matches over one year is consistently small for all institutions, ranging from 0.8% to 2%. This also significantly decreases when moving to the two year gap. In contrast, Scopus and MSA seem to have more varied results for the one year gap across institutions and with generally higher percentages than those of WoS.

The one standout case is ITB, an Indonesian university situated in the City of Bandung. Its results for WoS are similar to other institutions, but one-year comparisons from Scopus and MSA yielded disagreements on publications year for 24.6% and 25.6% of all common DOIs, respectively. We believe that this may be due to two reasons. Firstly, WoS has a significantly lower coverage of ITB (see Venn diagrams for ITB in Supplementary Material 3) than

37 This is the total number of DOIs that are jointly covered by the sources listed in each column title. The numbers here differ slightly from the first Venn diagram in Figure 5 because there exist a small number of DOIs in each source that had repeated entries but fall in different years. The numbers of such cases for WoS, Scopus, and MSA are 1, 2, and 43 respectively.

38 See Harzing and Alakangas (2017b), Hug and Brändle (2017), and https://academic.microsoft.com/faq.

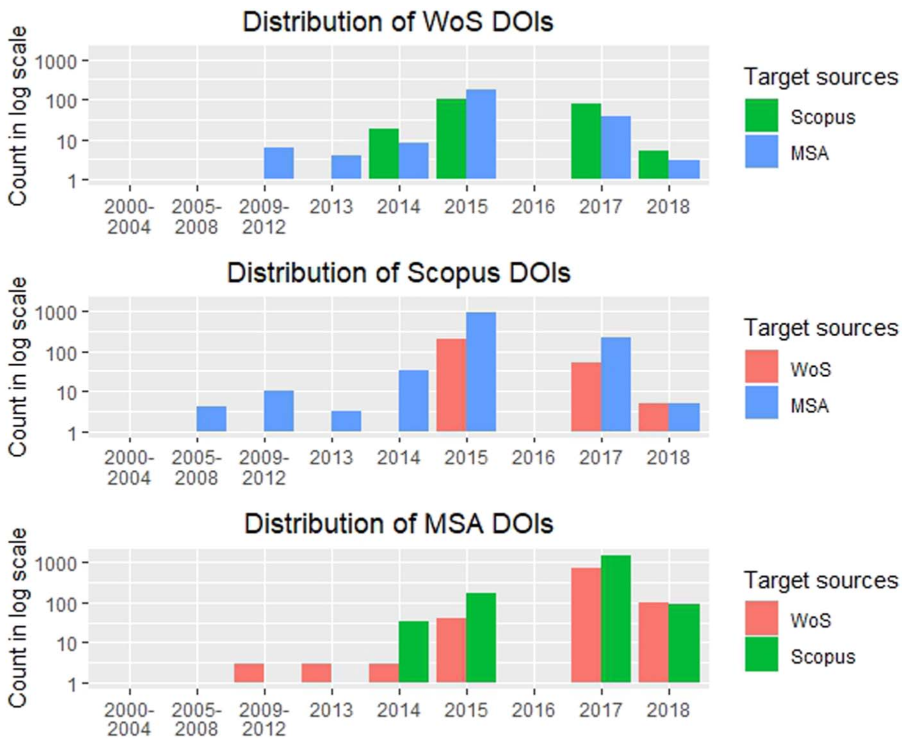

Figure 11. Number (log scale) of 2016 DOIs from each source (exclusively) that falls in another source but in a different year (15 universities combined).

Scopus and MSA. There is also a much lower number of DOIs exclusively indexed by WoS. Secondly, Indonesia has an extraordinarily large number of local journals owned by universities, and many of these are open access. This is largely driven by government policy, which requires academics and students to publish research results and theses in academic journals[39]. Many of these journals are also linked to conference output. This may have resulted in a systematic difference on how publication years (or dates) are recorded (or defined). The other two cases that stand out, although less extreme, are Cairo and IISC.

In Supplementary Material 7, the directions of the comparisons are displayed in more detail for the three standout cases (i.e., Cairo, IISC, and ITB). The comparisons are also narrowed down to just Scopus and MSA. It is immediately clear that the differences between Scopus and MSA are the main contributors to these standout cases. Also, it appears that MSA tends to record the publication year one year earlier than Scopus. This is in line with our earlier comments regarding MSA recording the date of first online publication and the publishing venues in Indonesia.

Let us now focus on the outer parts of the Venn diagrams (i.e., DOIs that appear to be exclusively indexed by a single source). The results for these sets of DOIs are presented in Table 7. Columns 2, 5, and 8 lists the number of 2016 DOIs exclusively indexed by WoS, Scopus, and MSA (compare these again with Venn diagrams in Supplementary Material 3), without checking against DOIs listed in other years. The subsequent columns list the

[39] See for example: https://www.openaire.eu/blogs/open-science-in-indonesia and https://campuspress.yale.edu/tribune/creating-an-open-access-indonesia/.

Table 6. Percentage of 2016 DOIs[40], per bibliographic source, listed in the other two sources but a year away (i.e., 2015 and 2017) and two years away (i.e., 2014 and 2018)

	All DOIs from WoS vs other two sources		All DOIs from Scopus vs other two sources		All DOIs from MSA vs other two sources	
Institution	**1 year**	**2 years**	**1 year**	**2 years**	**1 year**	**2 years**
Cairo	1.5	0.2	6.0	0.2	5.4	0.4
Curtin	1.2	0.2	2.3	0.0	2.6	0.3
DUT	1.3	0.1	2.5	0.1	3.1	0.3
IISC	1.3	0.2	4.9	0.2	6.0	0.1
ITB	1.0	0.0	24.6	0.0	25.6	0.0
LU	1.6	0.1	2.7	0.1	3.4	0.3
MIT	1.2	0.2	1.6	0.1	2.6	0.4
MSU	0.8	0.1	0.7	0.1	1.3	0.1
UNAM	1.5	0.1	1.5	0.0	1.7	0.2
UCL	0.8	0.1	1.2	0.2	2.3	0.3
UCT	1.8	0.1	2.0	0.1	2.8	0.1
Giessen	1.1	0.1	1.4	0.4	1.8	0.1
USP	1.5	0.1	2.2	0.1	2.3	0.2
Tokyo	1.2	0.2	2.1	0.1	2.7	0.3
WSU	2.0	0.2	2.4	0.2	2.1	0.2

percentages of these DOIs that can be matched against DOIs in other sources for one-year and two-year gaps, respectively. Consistent with Table 6, significantly higher portions of DOI matches occur after incorporating the first one-year gap, as compared to including a further one year on both sides. ITB again sees the largest impact of these inconsistencies, which corresponds to the observation made in Table 6. In general, the effect on these exclusive sets of DOIs varies considerably across institutions and sources (more so than observed in Table 6, as expected).

4.3. Document Types

Another important bibliographic variable is the document type (e.g., journal articles, proceedings, book chapters) that relates to each DOI. In particular, the coverage of different document types can lead to insights into potential disciplinary biases in data sources and differences in institutional focuses on output types.

For this study, we use the "genre" variable in Unpaywall metadata to determine the document type of each DOI. These are Crossref-reported types for all DOI objects in the Crossref

[40] Calculated out of all DOIs, from the particular source.

Table 7. Percentages of 2016 DOIs exclusively from each bibliographic source that is indexed by other sources within 1 and 2 years (before and after 2016) gaps, respectively

	DOIs excl. from WoS			DOIs excl. from Scopus			DOIs excl. from MSA		
Institution	**Original**	**1 year**[41]	**2 year**[42]	**Original**	**1 year**	**2 year**	**Original**	**1 year**	**2 year**
Cairo	340	2.4	0.3	261	47.9	0.0	660	21.2	1.8
Curtin	220	5.9	2.3	198	28.8	0.5	502	13.3	1.4
DUT	187	8.6	0.5	794	9.4	0.0	533	18.4	1.9
IISC	126	6.3	0.8	177	49.2	0.0	712	19.2	0.3
ITB	61	3.3	0.0	410	62.4	0.0	545	47.5	0.0
LU	127	4.7	0.0	138	18.8	0.0	307	15.0	1.6
MIT	1,309	2.8	0.3	396	20.2	0.8	1,784	8.8	1.6
MSU	530	3.0	0.2	206	8.3	0.5	1,148	5.7	0.6
UNAM	522	6.5	0.4	460	8.7	0.2	1,532	5.2	0.8
UCL	2,680	1.7	0.2	1,193	6.8	0.8	1,735	11.4	1.4
UCT	252	6.0	0.4	202	12.9	1.0	681	11.6	0.3
Giessen	234	2.6	0.0	110	11.8	3.6	303	7.6	0.7
USP	1,258	4.2	0.5	932	15.9	0.5	4067	7.1	0.6
Tokyo	732	8.9	0.1	565	26.7	1.4	1,793	13.2	0.8
WSU	265	7.9	1.1	122	23.0	2.5	595	8.1	0.8

database[43]. Table 8 provides the counts of each document type within each part of the Venn diagram between WoS, Scopus, and MSA (for all 15 institutions from 2000 to 2018 combined)[44]. An immediate observation is that journal articles make up (by far) the highest proportion of the DOIs. This is true overall and for individual parts of the Venn diagram, as would be expected. The scenario is again more interesting when we consider the outer parts of the Venn diagram (sections W, S, and M of the Venn diagram). The set of DOIs exclusive to MSA contains significantly more book chapters and proceeding papers relative to any other parts. It also provides almost all thesis entries in our data and is the only source to provide posted content (i.e., web pages and blogs). On the other hand, Scopus seems to provide many books and monographs not indexed by the other two sources.

Again we would like to examine how the situation plays out for individual institutions. After filtering the sets of DOIs to each institution and to the year 2016, we follow the same

[41] Percentage of DOIs from WoS only that are also indexed by at least one of the two other sources but recorded a year apart (in both directions).

[42] Percentage of DOIs from WoS only that are also indexed by at least one of the two other sources but recorded two years apart (in both directions).

[43] See https://unpaywall.org/data-format.

[44] Note here that the total number of DOIs are slightly lower in each part of the Venn diagram as compared to the left Venn diagram in Figure 5. This is because here we are only including DOIs that are also recorded in Unpaywall.

Table 8. Document types of all DOIs, recorded in Unpaywall, for all 15 universities combined from 2000 to 2018

	All 15 universities combined						
Section of Venn diagram[45]	WSM	SM	WM	WS	M	S	W
book-chapter	241	5,745	4,699	50	37,138	6,709	1,523
journal-article	393,524	96,600	95,127	144,634	190,768	63,244	85,850
proceedings-article	1,849	12,172	6,201	1,504	61,546	4,766	3,270
reference-entry	191	96	78	89	1,107	122	252
report	1	2	0	0	216	8	7
book	0	272	0	0	514	4,189	29
component	0	2	0	2	106	59	37
journal	0	7	0	1	9	8	0
journal-issue	0	1	3	9	200	20	26
monograph	0	64	0	0	144	505	1
other	0	14	1	0	36	220	64
dataset	0	0	0	0	7	0	2
dissertation	0	0	0	0	341	5	0
posted-content	0	0	0	0	1,391	0	0
reference-book	0	0	0	0	1	26	1
report-series	0	0	0	0	116	0	0
book-section	0	0	0	0	0	4	0
book-set	0	0	0	0	0	3	0
proceedings	0	0	0	0	0	0	1

procedure as above to produce the spread of document types across each part of an institution's Venn diagram. These are recorded in Supplementary Material 8. As we have observed for the combined data set, journal articles make up the highest portion of the DOIs for each institution. The next two most common document types are book chapters and proceeding papers. The only exception is ITB, where there are slightly more proceeding papers than journal articles. Interestingly, there are a few universities with more book chapters than proceeding papers (Curtin, UNAM, UCL, UCT, Giessen, and WSU).

There are high proportions of book chapters indexed exclusively by MSA for all institutions. MSA also has the highest proportion of exclusively indexed journal articles, except for MIT, UCL, and Giessen (WoS has the highest such proportion for these three institutions). It is also observed that MSA and Scopus seem to bring in more additional proceeding papers than WoS (the only exception being UNAM, where all three sources have similar exclusive coverage on proceeding papers). Scopus also seems to often add books and monographs not indexed by

[45] See Figure 3 for the labelling of the Venn diagram.

Table 9. Total citations for all 15 institutions from 2000 to 2018, as per OpenCitations

	WoS	Scopus	MSA	Combined
Number of DOIs[46]	735,832	734,515	907,239	1,202,032
Total citations[47]	9,670,953	9,581,710	9,122,420	13,060,486
Citations per output	13.1	13.0	10.1	10.9

the other two sources. For all universities, journal articles make up the majority of DOIs exclusively indexed by WoS. In contrast, the document types of DOIs exclusively indexed by Scopus or MSA are more diverse. Overall, we observe that each source has a different exclusive coverage of document types, and this coverage also varies across institutions.

4.4. Citation Counts

One set of commonly used bibliographic metrics in the evaluation of academic output are those that relate to citation counts. These include metrics such as *h*-index, impact factor, and eigenfactor. However, these citation metrics can also be calculated via different sources. WoS, Scopus, and MSA all record and maintain their own citation data. While some research has shown that the citation counts across these sources showed high correlations at the author level and journal level (Harzing, 2016; Harzing & Alakangas, 2017a, 2017b), the corresponding effects on a set of universities remain relatively unknown. We match each DOI against the list of DOI citation links in OpenCitations' COCI data and obtain (if it exists) its total citation count. In Table 9, we present the results combining DOIs for our initial set of 15 universities for all years from 2000 to 2018.

To repeat, our goal is to identify the effects of coverage on the discoverability of sets of outputs that would then be evaluated using an external source of data. For this reason we use the COCI data from OpenCitations to provide an external and independent source of data that can be comprehensively applied to the evaluations of each set of DOIs that we discover for each institution and year. Similarly, when we use the same sets of outputs to compare OA status (section 4.5) we use an external data source (Unpaywall) that allows us to evaluate the performance of the sets of identified outputs. This is different from comparing the results of using each bibliographic data source as the source of both the sets of outputs and the source of performance data. For completeness, we also provide such an analysis in Supplementary Material 9. In the event, the conclusions of both analyses are very similar. This suggests that COCI is a viable source of comprehensive citation data for cross comparison at system level, even if it is not an appropriate source of data in its current form for analyzing the comparative performance of individual outputs due to its lack of coverage of some sources of citations.

The results show that the total number of citations to MSA DOIs is slightly lower than in WoS and Scopus. This is in addition to an already larger set of (Unpaywall/Crossref) DOIs. Hence, MSA resulted in a lower average citation number (from a smaller numerator of citation counts and a larger denominator of Crossref DOIs) from the OpenCitations citation links.

[46] This is the number of DOIs in each source that is also indexed by Unpaywall (i.e., Crossref).

[47] These are calculated using the sets of DOIs from each source that are also indexed in Unpaywall. OpenCitations and Unpaywall both use Crossref DOIs as identifiers. If we use the full set of DOIs (i.e., including non-Crossref DOIs), we get a very small increase in citation totals, ranging from 0.01% to 0.03%.

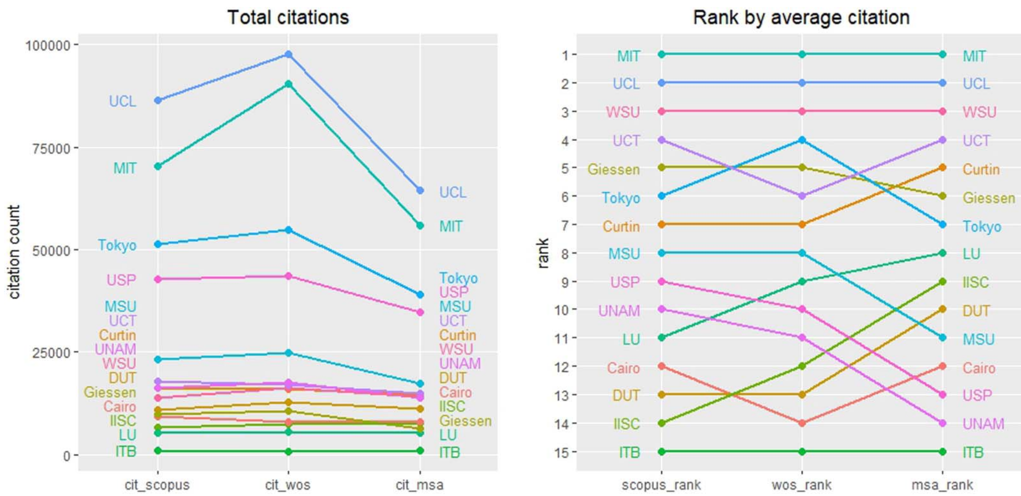

Figure 12. Total citations and rankings in average citations for 15 institutions (in 2016), as per bibliographic source[48].

As a further analysis, we would like to investigate how the change of bibliographic source influences the perceived performance of an institution while holding the evaluative data source constant. Figure 12 presents two different charts for total citations and ranks by average citations for each of the sample of 15 universities. UCL and MIT experience the biggest changes in total citation counts: decreases of 34% and 38%, respectively (left-hand chart in Figure 12), when shifted from WoS to MSA. While the remaining universities' total citation counts seem to have changed to a lesser degree across sources, the differing coverage of DOIs (i.e., different number of DOIs recorded) by each source can still significantly change the average citation counts. This is evidenced in the second chart of Figure 12. Only four universities' rankings remain unchanged across sources (the top three and last place). All other universities' positions have shifted at least once across the three sources, with the biggest changes affecting IISC, USP, and UNAM.

For a more general view, we now include the ranking results for the larger set of 155 universities in Figure 13. The results related to universities that have shifted by at least 20 positions (at least once) across the three sources are highlighted in color, with universities from English-speaking regions in red and non-English-speaking ones in orange. This includes 45 universities: 27 in red and 18 in orange. That means that almost one-third of the universities have shifted 20 or more positions. The most extreme cases include Charles Sturt University dropping 146 places when moved from WoS to Scopus, and Universitat Siegen and University of Marrakech Cadi Ayyad dropping 143 and 112 positions, respectively, when moved from WoS to MSA.

For further insight into the distribution of shifts across sources, we summarize the pairwise changes to average citations and rankings by average citations into box plots in Figure 14. The median change to average citations when moving from WoS to Scopus is just below zero, while the corresponding medians for WoS to MSA and Scopus to MSA are both just above zero. The corresponding mean values are −0.2, 1.2, and 1.3, respectively. As for the changes to rankings, the median and mean values are all close to zero. The distributions of these box plots are characterized by a concentrated center with long tails. This signifies the existence of two contrasting groups: those universities that were less affected by shifts in bibliographic

[48] Only Unpaywall (i.e., Crossref) DOIs are included in the calculations of average citations.

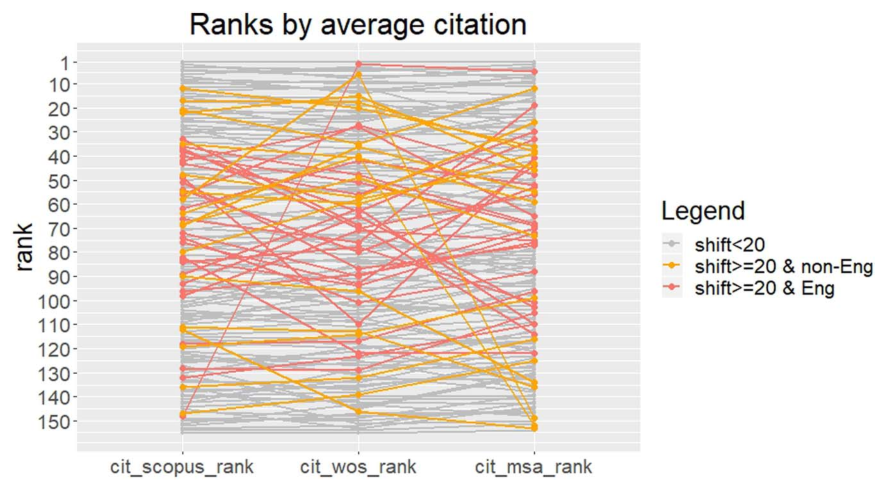

Figure 13. 2016 ranking by average citations for 155 universities, as per bibliographic source (with those shifting at least 20 positions displayed in color).

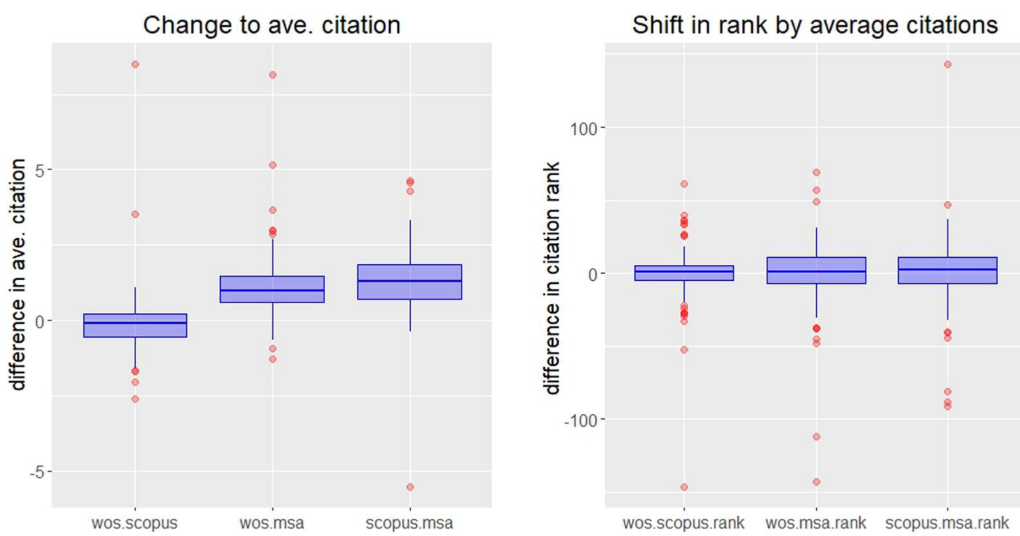

Figure 14. Changes to 2016 average citations (left) and rank by average citations (right) when moving from one source to another for 155 universities.

sources, and those that can have their performance levels, in terms of average citations, greatly altered depending on the choice of source.

4.5. Open Access Status

A recent topic of interest is the number of OA publications produced at different levels of the academic system. In particular, universities may wish to evaluate their OA standings for compliance with funder policies and OA initiatives. For objects with DOIs (and, in particular, Crossref DOIs), various information on accessibility can be queried through Unpaywall[49]. We matched all DOIs from the sample of 15 universities to the Unpaywall metadata and

[49] https://unpaywall.org/.

Table 10. Total level of OA for all DOIs in our sample of 15 universities, from 2000 to 2018, as per bibliographic source

	WoS	Scopus	MSA	Combined
Number of DOIs[50]	735,832	734,515	907,239	1,202,032
OA count	317,021	294,655	367,100	498,929
%OA	43.1	40.1	40.5	41.5

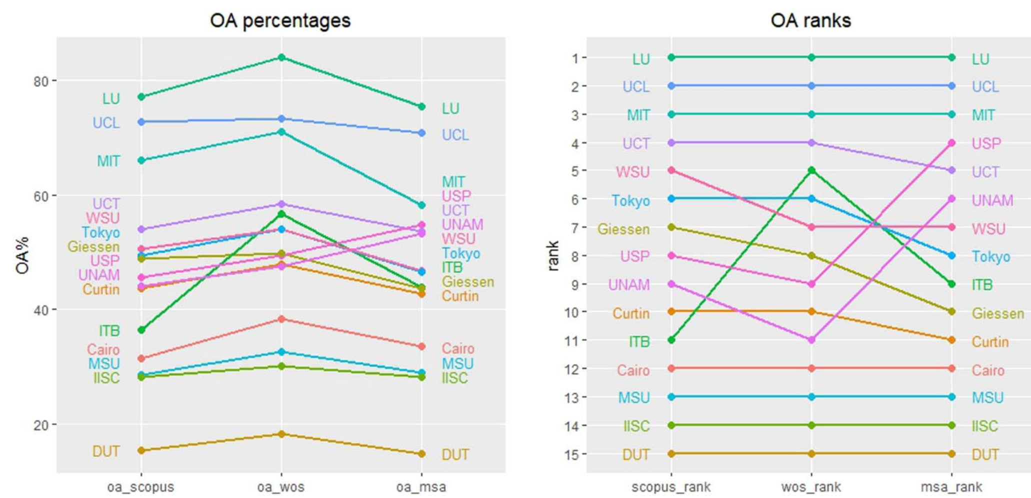

Figure 15. 2016 Total OA percentages (left) and OA rankings (right) for 15 institutions, as per bibliographic source[51].

calculated the percentage of OA output across each bibliographic source and for all (unique) DOIs combined. This is presented in Table 10.

There do not appear to be substantial changes to the overall OA percentage when shifting across sources for the combined sets of DOIs. However, we should keep in mind that there are significant differences in each source's DOI coverage, as observed earlier.

To see whether such consistency in OA percentages carries over to the institutional level for 2016, we again filter the data down to each university. Figure 15 provides the percentages of OA output and the corresponding relative ranks for each institution, as per the set of 2016 DOIs indexed by each source and also recorded in Unpaywall. It is observed that, for quite a few universities, the OA percentages vary considerably depending on which source is used to obtain the sets of DOIs. The most extreme case is again ITB, which had about a 20% drop when moving from WoS to Scopus. Also, the direction of OA percentage changes differs across universities. For example, the OA percentage for MIT decreased when moving from Scopus to MSA, but the opposite occurred for USP. This is especially critical if one is to compare the relative OA status across universities, which can vary according to the source of DOIs used. As for OA ranks, it seems to indicate a group of universities not affected by changing source, while the other group have their ranks shifted significantly. The most affected cases seem to be USP, ITB, and UNAM.

[50] This is the total number of DOIs in each source that are recorded in Unpaywall.

[51] Only DOIs indexed by Unpaywall are included in the calculations.

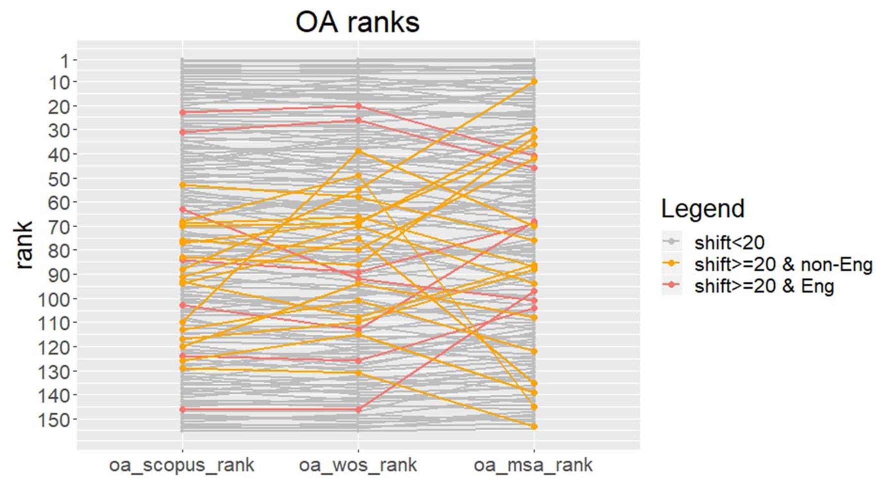

Figure 16. 2016 OA rankings for 155 universities, as per bibliographic source (with those shifting at least 20 positions displayed in color).

The effects on OA levels and ranks are more difficult to express directly for the larger set of 155 universities. Again, instead of labelling the full set of universities, we highlight only those that have shifted by 20 positions or more at least once. This is displayed in Figure 16. There are 24 out of 155 universities that have shifted at least 20 positions in OA ranking when moved across sources. Seventeen of these are from non-English-speaking regions, including six Latin American universities (out of seven in the full set). This is an indication of the potential difference in coverage of the three sources due to language.

Analogous to the earlier analysis on citations, we calculate differences in OA percentages and OA ranks when shifting from one source to another and present these in a number of box plots in Figure 17.

Evidently, the median OA percentage changes when shifting from WoS to Scopus, WoS to MSA, and Scopus to MSA are all positive. The corresponding mean changes are also positive

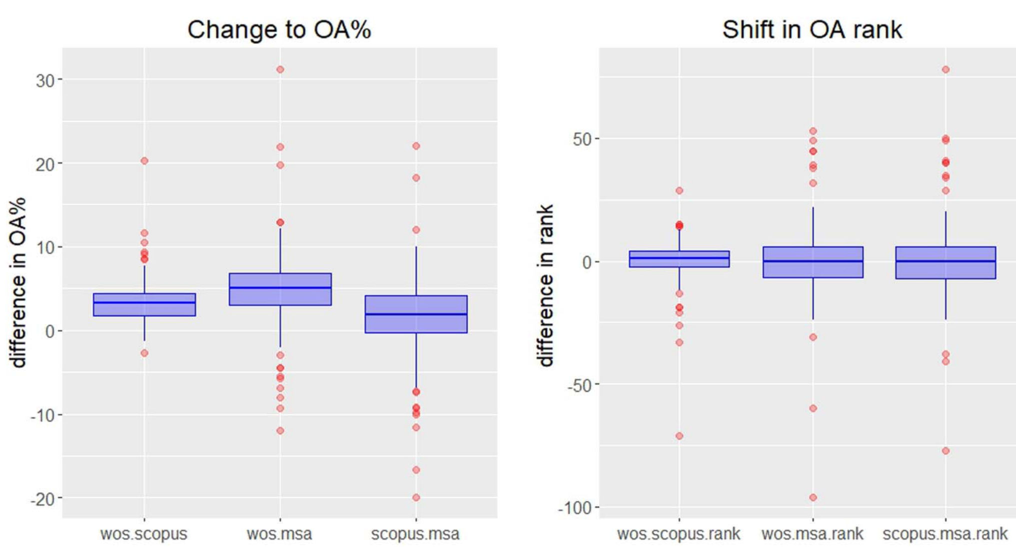

Figure 17. Changes to 2016 OA percentage (left) and OA rank (right) when moving from one source to another for 155 universities.

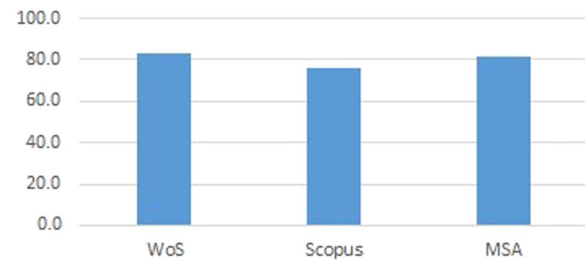

Figure 18. Percentage of DOIs with plausible affiliation as per matching against original document or the other two sources.

at 3.4%, 4.9%, and 1.5% respectively. The median and mean changes to rankings are all close to zero. However, in both OA percentage and OA rank changes, there are many recorded extreme points (including both negative ones and positive ones). These include an OA percentage change as large as 31.1% (moving from WoS to MSA) and an extreme drop in OA rank of 96 positions (MSA to WoS). The general distributions of both changes to OA percentage and changes to OA rankings are characterized by high central peaks and long tails. This implies that, while the changes are small for most of the universities, there are also a significant number of cases where universities are largely affected by shifts in data sources.

4.6. Manual Cross-Validation

This section provides a summary of our manual cross-validation results of DOIs exclusively indexed by each source. For each of the 15 institutions, we randomly sampled 40, 30, and 30 DOIs from their sets of 2016 DOIs exclusively indexed by WoS, Scopus, and MSA, respectively (i.e., sections W, S, and M from the Venn diagram in Figure 4). This was done after the removal of DOIs that match up to other sources in a different year (this includes the neighboring two years: 2014, 2015, 2017, and 2018). Subsequently, these lists of DOIs go through a thorough manual cross-validation process. Various questions were asked against each DOI and compared across the three bibliographic sources. These are summarized in a table in Supplementary Material 10.

In the following, we shall highlight some of the main findings in a few simple charts, with further detailed analysis provided in Supplementary Material 10. Firstly, we focus on the plausibility of affiliation associated with each DOI.

In Figure 18, we present results related to affiliation of each DOI as per source. For each DOI, the target affiliation is checked against its online original document[52]. When the original document is not accessible (e.g., not OA), the affiliation is matched against the other two sources. The decision is made to indicate the affiliation as plausible when the target affiliation (i.e., affiliation as per our data collection process) appears exact (including obvious versions of the university name) on the document, a plausible affiliation name variant[53] appears on the document, or the affiliation is confirmed by at least one of the other two bibliographic sources. This should (roughly) inform us about whether each source has correctly assigned these DOIs to the target affiliations.

[52] This is done via doi.org as a first pass, followed by a manual title search online.

[53] The decision of whether an affiliation is a plausible variant of the target affiliation is made somewhat subjectively, but informed via simple online searches. These may include subdivisions under the target affiliation (e.g., departments, research groups), aliases, etc. The strategy is that this should be a simple decision via a quick online search; otherwise a negative response is recorded.

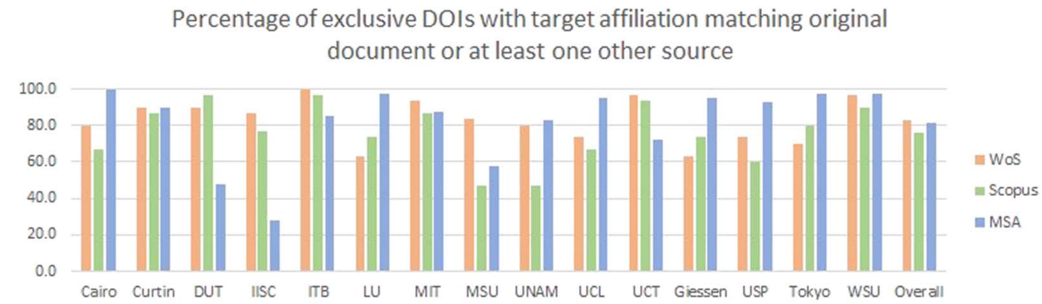

Figure 19. Percentage of exclusive DOIs from one source that has a plausible match to target affiliation as per original document or at least one other source.

The result shows that all sources have only correctly assigned roughly 80% of their respective DOIs from our sample to the target affiliations, with very little difference in performance across the sources. When this is filtered down by university (see Figure 19), we see a more varied performance across universities.

Interestingly, not all percentages are very high across the universities. This is especially apparent for DUT and IISC, where MSA seems to have affiliated many DOIs to these two institutions without the target affiliations actually appearing on the original documents or confirmed by another source. Similarly, for DOIs that were assigned to MSU and UNAM by Scopus alone, only 46.7% (for both institutions) have a plausible affiliation match.

We have also checked each DOI against the DOI string actually recorded as per original document (where applicable) or via doi.org. These percentages (of correct DOIs) are 93.1%, 98.2% and 96.7 for WoS, Scopus, and MSA respectively (with all 15 institutions combined). While these numbers are relatively high, a significant number of errors suggests that DOIs are not being systematically checked against authoritative sources such as Crossref which we find surprising. In addition the nature of these errors which in some cases appear to be transcription or OCR errors is concerning (see examples in Supplementary Material 11).

We now take an overview of results from the DOI and title matching, given in Figure 20. As an initial analysis, no affiliation information is considered here and the results represent all DOIs for the 15 universities combined. Each bar represents the percentage of output corresponding to DOIs (that initially appear to be) exclusively indexed by one source that can

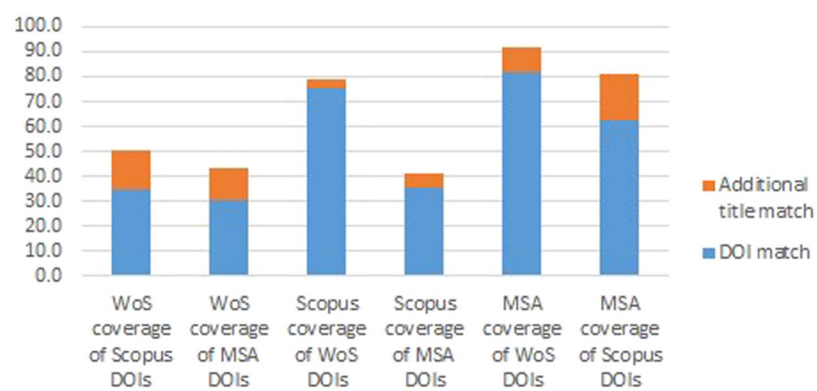

Figure 20. Percentage of DOIs found in another source by DOI and title matching.

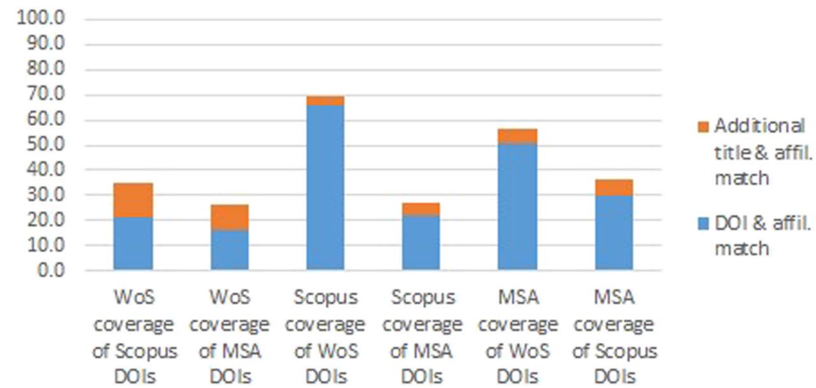

Figure 21. Percentage of DOIs found in another source by DOI and title matching, combined with affiliation matching.

be found in another source by DOI matching and title matching (via manual searches online). For example, the first bar corresponds to objects with DOIs sampled from Scopus. The height of the blue bar shows the percentage of these objects that can be found in WoS by DOI matching. The orange bar then indicates how much more can be found by title matching.

We found that in all cases where there is a DOI match, there is also a title match. However, the opposite is not necessarily true. Hence, title matching increases the coverage slightly in all scenarios. This does imply that all three sources have missing DOIs in their metadata, though there appear to be fewer cases for Scopus. Scopus also seems to have a good coverage of DOIs from WoS. More strikingly, a very high proportion of DOIs and titles from WoS and Scopus are found in MSA. In contrast, far fewer MSA DOIs and titles are covered by WoS and Scopus.

In Figure 21, we added affiliation matching to the mix; that is, we checked whether the target affiliation (i.e., affiliation as per our data collection process) appears in the metadata of the matching source after an object is found by DOI or title match. This decreased the coverage in all cases, indicating the potential disagreement of affiliation across sources. MSA is the most affected of the three sources.

The general picture that has emerged is that MSA seems to have good coverage of DOIs that initially appeared to be exclusively from WoS or Scopus. However, it falls short on correctly assigning affiliations and recording DOIs corresponding to each output. MSA also seems to have substantially broader coverage, including many objects that genuinely appear exclusively in MSA. The correctness of affiliation metadata for these is high overall, but tends to vary across institutions.

5. LIMITATIONS AND CHALLENGES

One obvious limitation is our focus on DOIs and our dependence on the uniqueness of DOIs. We do note that there may be research objects with multiple DOIs and related objects may also be assigned a common DOI (e.g., books can fit both cases). A related matter is the correctness of DOIs; that is, whether they were recorded correctly (as per doi.org) in each source's metadata[54]. DOIs that did not generate Unpaywall returns could include such cases. While our manual cross-validation process did check our samples against doi.org, it is not clear what the scale of this issue is for the overall data.

[54] See Gorraiz, Melero-Fuentes, et al. (2016) for a discussion of the availability of DOIs in WoS and Scopus.

Our manual cross-validation process was carried out over a number of months after the initial data collection process. This means that there may be potential discrepancies between metadata content at the time of collection and time of manual search. However, we expect such cases to be few, given that we are focused on 2016 data, and a number of manual spot checks did not reveal any obvious such cases.

Both API and manual searches for WoS and Scopus may be limited to the subscription model of the authors' home institution at the time of access. On the other hand, matching identifiers have also proved to be challenging. For example, a few institutions have multiple Scopus IDs (e.g., multiple campuses), without an overarching ID. For the three cases we have encountered, out of the 155 universities we have selected what appeared to be the main campus IDs. A recent study (Donner, Rimmert, & Van Eck, 2020) have also highlighted issues regarding the affiliation disambiguation systems in WoS and Scopus. Other challenges include Unpaywall and OpenCitations coverage is limited to Crossref DOIs, manual cross-validation is limited to DOI and title searches only, and there is inherent subjectivity in linking plausible affiliation names.

6. CONCLUSION

This article has taken on the task of comparing and cross-validating various bibliographic characteristics (including coverage, publication date, OA status, document type, citations, and affiliation) across three major research output indexing databases: WoS, Scopus, and MSA. This is done mainly with a focus on identifying institutional-level differences and the corresponding effects of using different data sources in comparing institutions. Our data consist of all objects with DOIs extracted from the three bibliographic sources for an initial sample of 15 universities and a further supplementary 140 universities (used only where applicable).

Firstly, we found that the coverage of DOIs not only differs across the three sources, but their relative coverages are also nonsymmetrical, and the distribution of DOIs across the sources varied from institution to institution. This means that the sole use of one bibliographic source can potentially seriously disadvantage some institutions and advantage others in terms of total output. While the general level of agreement on publication year is high across sources, there were individual universities with large differences in coverage per year. The comparison of document types showed that different sources can systematically add coverage of selected research output types. This may be of importance when considering the coverage of different research discipline areas.

Our subsequent analyses further showed that while the aggregate levels (i.e., for 15 universities combined) in citation counts and OA levels varied little across sources, there are significant impacts at the institutional level. There were clear examples of universities shifting dramatically in both of these metrics when moving across sources, some in opposite directions. This makes any rank comparison of citations or OA levels strongly dependent on the selection of bibliographic source.

Finally, we implemented a manual cross-validation process to check metadata records for samples of DOIs that initially appeared to be exclusive from each source, for each of the 15 universities. The records were compared across the three bibliographic sources and against (where accessible) the corresponding online research documents. The process revealed cases of missing links between metadata and search functionalities within each database (for both affiliation and DOI). This means that the real coverage of each source is unnecessarily truncated. Overall, it appears that MSA has the highest coverage of objects that initially appeared exclusive to other sources. However, it often has missing DOIs and affiliations that do not match with WoS, Scopus, or online documents.

There is also strong evidence that the effects of shifting sources may be more prominent for non-English-speaking and non-European universities. Similar signs were observable for universities that are low ranked or medium ranked in both citations and OA levels, while those that achieve high rankings in these measures show much smaller shifts in position when the data source is changed. Universities that are highly ranked on these measures also tend to be highly ranked in general rankings like the THES, suggesting a bias in reliability and therefore curation effort toward prestigious universities.

Our concluding message is: Any institutional evaluation framework that is serious about coverage should consider incorporating multiple bibliographic sources. The challenge is in concatenating unstandardized data infrastructures that do not necessarily agree with each other. For example, one primary task would be to standardize the publication dates, especially for longitudinal study. This may be possible, to a certain degree, by using Crossref or Unpaywall metadata as an external reference set. The development of the Research Organization Registry[55] may also provide further opportunities to improve on disambiguation of institution names via a community-managed data source. Tackling these problems is by no means trivial. However, it has the potential to greatly enhance the delivery of fairer and more robust evaluation.

ACKNOWLEDGMENTS

The authors would like to thank Alberto Martín-Martín and an anonymous reviewer for their valuable feedback, which has helped to improve this article.

AUTHOR CONTRIBUTIONS

Chun-Kai (Karl) Huang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. Cameron Neylon: Conceptualization, Data curation, Investigation, Methodology, Software, Project administration, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. Chloe Brookes-Kenworthy: Conceptualization, Data curation, Investigation, Writing—review & editing. Richard Hosking: Conceptualization, Data curation, Investigation, Software, Writing—review & editing. Lucy Montgomery: Conceptualization, Project administration, Supervision, Writing—review & editing. Katie Wilson: Conceptualization, Writing—review & editing. Alkim Ozaygen: Conceptualization, Writing—review & editing.

COMPETING INTERESTS

The authors declare there to be no competing interests. The funder and internal university sponsor had no part in designing the study or describing the results.

FUNDING INFORMATION

This work was funded by the Research Office of Curtin University through a strategic grant, the Curtin University Faculty of Humanities, and the School of Media, Creative Arts and Social Enquiry.

[55] https://ror.org/.

DATA AVAILABILITY

The raw data collected from WoS and Scopus cannot be made available publicly due to the respective licensing terms. However, the curated and derived secondary data are made available, alongside the code used for the analyses, at Huang et al. (2019), as referenced in the main text. The data collected for the manual cross-validation (section 4.6) is fully available at Brookes-Kenworthy, Huang, et al. (2019), as referenced in the main text.

REFERENCES

Anderson, M. S., Ronning, E. A. , De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*(4), 437–461. https://doi.org/10.1007/s11948-007-9042-5

Archambault, E., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, *60*(7), 1320–1326. https://doi.org/10.1002/asi.21062

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(7), 1–8. https://doi.org/10.1186/1742-5581-3-7

Bar-Ilan, J. (2008). Which *h*-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271. https://doi.org/10.1007/s11192-008-0216-y

Brookes-Kenworthy, C., Huang, C.-K., Neylon, C., Wilson, K., Ozaygen, A., Montgomery, L., & Hosking, R. (2019). Manual cross-validation data for the article: "Comparison of bibliographic data sources: Implications for the robustness of university rankings" [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3379703

De Domenico, M., Omodei, E., & Arenas, A. (2016). Quantifying the diaspora of knowledge in the past century. *Applied Network Science*, *1*(15), 1–13. https://doi.org/10.1007/s41109-016-0017-9

Donner, P., Rimmert, C., & Van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies*, *1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Effendy, S., & Yap, R. H. C. (2017). Analysing trends in computer science research: A preliminary study using the Microsoft Academic Graph. *Proceedings of the 26th International Conference on World Wide Web Companion*, 1245–1250. https://doi.org/10.1145/3041021.3053064

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS One*, *5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, *10*(4), 933–953. https://doi.org/10.1016/j.joi.2016.07.003

Giles, C. L., Bollacker, K., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. DL'98 Digital Libraries, *3rd ACM Conference on Digital Libraries*, 89–98. https://doi.org/10.1145/276675.276685

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*(1), 177–214. https://doi.org/10.1007/s11192-018-2958-5

Harzing, A. W. (2016). Microsoft Academic (Search): A Phoenix arisen from the ashes? *Scientometrics*, *108*(3), 1637–1647. https://doi.org/10.1007/s11192-016-2026-y

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Harzing, A. W., & Alakangas, S. (2017a). Microsoft Academic: is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383. https://doi.org/10.1007/s11192-016-2185-x

Harzing, A. W., & Alakangas, S. (2017b). Microsoft Academic Is one year old: The phoenix is ready to leave the nest. *Scientometrics*, *112*(3), 1887–1894. https://doi.org/10.1007/s11192-017-2454-3

Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. *Higher Education Management and Policy*, *19*(2), 1–24. https://doi.org/10.1787/hemp-v19-art12-en

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft Academic Graph. *D-Lib Magazine*, *22*(9/10). https://doi.org/10.1045/september2016-herrmannova

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2019). Codes and data for the article: Comparison of bibliographic data sources: Implications for the robustness of university rankings. Zenodo. https://doi.org/10.5281/zenodo.3541520

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551–1571. https://doi.org/10.1007/s11192-017-2535-3

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with Microsoft Academic. *Scientometrics*, *111*(1), 371–378. https://doi.org/10.1007/s11192-017-2247-8

Jacsó, P. (2005). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547. https://www.jstor.org/stable/24110924

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and

Google Scholar for articles published in general medical journals. *Journal of American Medical Association*, *302*(10), 1092–1096. http://doi.org/10.1001/jama.2009.1307

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Moore, S., Neylon, C., Eve, M. P., O'Donnell, D. P., & Pattinson, D. (2017). "Excellence R Us": University research and the fetishisation of excellence. *Palgrave Communications*, *3*, 16105. https://doi.org/10.1057/palcomms.2016.105

Neylon, C., & Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLOS Biology*, *7*(11), e1000242. https://doi.org/10.1371/journal.pbio.1000242

Norlander, B., Li, P., & West, J. D. (2018). Estimating article influence scores for open access journals. *PeerJ Preprints*, *6*, e26586v1. https://doi.org/10.7287/peerj.preprints.26586v1

Paszcza, B. (2016). Comparison of Microsoft Academic (Graph) with Web of Science, Scopus and Google Scholar. Master's thesis. University of Southampton.

Portenoy, J., Hullman, J., & West, J. D. (2016). Leveraging citation networks to visualize scholarly influence over time. *Frontiers in Research Metrics and Analytics*, *2*, 8. https://doi.org/10.3389/frma.2017.00008

Portenoy, J., & West, J. D. (2017). Visualizing scholarly publications and citations to enhance author profiles. *Proceedings of the 26th International Conference on World Wide Web Companion*, 1279–1282. https://doi.org/10.1145/3041021.3053058

Ranjbar-Sahraei, B., van Eck, N. J., & de Jong, R. (2018). Accuracy of affiliation information in Microsoft Academic: Implications for institutional level research evaluation. *Proceedings of the 23rd International Conference on Science and Technology Indicators*, 1065–1067. https://openaccess.leidenuniv.nl/handle/1887/65339

Sandulescu, V., & Chiru, M. (2016). Predicting the future relevance of research institutions – The winning solution of the KDD Cup 2016. arXiv:1609.02728v1. https://arxiv.org/abs/1609.02728v1

Shin, J. C., & Toutkoushian, R. K. (2011). The past, present, and future of university rankings. In: Shin J., Toutkoushian R., & Teichler U. (Eds.), *University Rankings. The Changing Academy – The Changing Academic Profession in International Comparative Perspective*, vol. 3, 1–16. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1116-7_1

Stergiou, K., & Lessenich, S. (2014). On impact factors and university rankings from birth to boycott. *Ethics in Science and Environmental Politics*, *13*(2), 101–111. https://doi.org/10.3354/esep00141

Thelwall, M. (2018). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, *12*(1), 1–9. https://doi.org/10.1016/j.joi.2017.11.001

Tsay, M.-Y., Wu, T.-L., & Tseng, L.-L. (2017). Completeness and overlap in open access systems: Search engines, aggregate institutional repositories and physics-related open sources. *PLOS One*, *12*(12), e0189751. https://doi.org/10.1371/journal.pone.0189751

Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics*, *11*(3), 766–782. https://doi.org/10.1016/j.joi.2017.05.014

van Wessel, M. (2016). Evaluation by citation: Trends in publication behavior, evaluation criteria, and the strive for high impact publications. *Science and Engineering Ethics*, *22*(1), 199–225. https://doi.org/10.1007/s11948-015-9638-0

Wesley-Smith, I., Bergstrom, C. T., & West, J. D. (2016). Static ranking of scholarly papers using article-level eigenfactor (ALEF). arXiv:1606.08534v1. https://arxiv.org/abs/1606.08534v1

Yang, K., & Meho, L. I. (2006). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the Association for Information Science and Technology*, *43*(1), 1–15. https://doi.org/10.1002/meet.14504301185