Dimity Stephen / Stephan Stahlschmidt / Paul Donner

KB Quality Assurance at the macro-level: Comparing the current and previous WoS snapshots

Report on wos_b_2019 and wos_b_2020

Version: 20200925

DZHW
German Centre for Higher Education Research and Science Studies

Editor:
German Centre for Higher Education Research and Science Studies (DZHW) GmbH
Lange Laube 12 | 30159 Hannover | Germany | info@dzhw.eu |www.dzhw.eu
POB 2920 | 30029 Hannover | Germany
phone: +49 511 450670-0 | fax: +49 511 450670-960
Chairman of the Supervisory Board:
Ministerialdirigent Peter Greisler
Scientific Director:
Prof. Dr. Monika Jungbauer-Gans
Managing Director:
Karen Schlüter
Registration Court:
Amtsgericht Hannover | HRB 6489
VAT No.: DE291239300
September 2020

Contents

Motivation

The aim of the report is to identify any potential changes in data between or within database versions that may indicate quality issues. To do so it offers:

- a visual comparison
- between time-series over the last 10 years
- stemming from the current and previous KB database snapshots
- on several key indicators
- for national, sectoral and institutional entities.

The DZHW already conducts quality assurance testing at the micro-level for KB bibliometric databases before the tables enter the production environment. This testing is invaluable to ensuring tables and variables contain the expected content. This report supplements the current micro-level approach by examining changes at the macro-level - institutions, sectors, countries, disciplines - in key variables between the latest two iterations of the databases.

This report is not an exhaustive analysis of the databases' content, nor does it investigate any anomalies identified within the databases. However, this report probes the core variables fundamental to common bibliometric analyses, serves as an overview of the current state of the databases, and highlights changes that may indicate issues with data quality that warrant further investigation to understand or rectify. Changes may arise through several means. For instance, the database provider may add or remove journals from indices, change the discipline classification, or change how the classification is applied. The KB may identify new or decommissioned institutions, which can affect publication output for particular disciplines, or countries may implement policies regarding publication practices that can exert a substantial influence on the content published over time. This report aims to provide users of the KB databases with an overview of potential changes soon after the databases enter the production environment, allowing these factors to be considered in analyses.

Set of indicators

The indicators we have chosen reflect the core variables in the database that are fundamental to key bibliometric analyses and indicators. We provide context to the selection of variables and what information can be determined from their analysis in each of the following sections.

We make two sets of comparisons in this report. For indicators where it is important to consider trends over time, such as whole publication counts, we compare the databases for the 10 years up to the year for which both have complete data. For example, the latest common year with complete data for the wos_b_2019 and wos_b_2020 databases is 2018, as data for the absolute latest year in each database are incomplete. Similarly, where citation-based indicators are used, we present the time-series up to the latest common year with complete citation data, which is 2016 for the wos_b_2019 and wos_b_2020 databases. This comparison highlights any differences in trends between the databases for the most recent decade.

For other indicators, it is most useful to compare changes between just the most recent years of complete data in each database. For instance, we examine the threshold for Excellence Rates in 2016 from the wos_b_2019 database against 2017 in the wos_b_2020 database. Changes between the years are expected given we are comparing two different sets of publications, however this comparison can also provide insight into structural changes between the database iterations, such as the addition or removal of journals from indices, which may influence indicators at the macro-level.

Such comparisons are also helpful in identifying new or removed institutions or discipline categories. Further, although users will likely use the latest database to produce a complete time-series for new analyses, it is important to understand how additional years of a time-series might differ to existing time-series presented in publications and reports.

Set of entities

We have chosen to compare the databases at the national, sectoral, and institutional levels. The countries chosen are based on those most commonly examined by the DZHW due to their status as high-performing countries or as countries against which it is useful and informative to compare Germany.

We also examine the key German sectors: Universities (Uni), Fachhochschulen (FH), Max Planck Gesellschaft (MPG), Fraunhofer Gesellschaft (FHG), Helmholtz Gemeinschaft (HGF), Leibniz Gemeinschaft (WGL), the business sector (Econ), non-university hospitals (Klinik), and combined Ressortforschung-Bund and Ressortforschung-Laender (Gov). The remaining smaller sectors, such as research associations, clubs, and international and foreign organisations are grouped into an "other" category. Individual institutions are also examined, however only for Germany due to the unavailability of institutional coding for other countries. Further, given the large number of institutions, we present only the institutions that appear to have suddenly stopped or started publishing, and the larger institutions that have shown substantial changes in the indicator of interest.

Methodological details

Please note the following methodological details. First, we focus on articles and reviews published in journals as these are the most common documents used in bibliometric analyses. As previously noted, we supply a shortened time-series for citation-based indicators to allow for a 3-year citation window. Wang [2] determined that at least 3 years is required for publications to reach their maximum number of citations per year, after which point the number of citations are likely representative of the publication's long-term impact. As such, citation-based indicators include all citations received within the publication year and the subsequent two years.

Whole counting is used throughout the report. Although it is most common to use fractional counting, analysing variables using whole counts will still reveal potential changes in the variables, negating the need to spend the additional time required to set up the necessary tables to perform fractional counting before this report can be run.

Data for disciplines are presented based on either the sc_traditional, sc_extended or Research Areas (RA) classification. Sc_traditional is the fine-grain classification more commonly used in analyses by the DZHW. However, as it contains over 250 categories, it is sometimes useful to use a coarse-grain approach to present an overview of the disciplines. As such, we present some data on the RA classification, which collapses the disciplines into five broad groups. RA are based on the sc_extended classification and so, as Clarivate only provides mapping between the RA and sc_extended classifications, supplementary tables presenting underlying data for RA are presented using the sc_extended classification. Each section containing data about disciplines notes which classification is used.

This report is automated and so tables are created regardless of whether any data fit the criteria, as such blank tables may appear in this report and are nonetheless informative about the indicator under examination.

Analysis

Whole count of articles and reviews: Selected countries and German sectors

The count of items produced by selected entities is the most fundamental bibliometric indicator. Given publication counts form the basis of many indicators, understanding the time-series trend within and between databases can inform expectations about potential changes that may arise in other indicators. In Figures 1 and 2 we present the whole counts of articles and reviews published in journals over the last 10 years by selected countries and sectors. Please note that the panels have different axes.

Changes in publication counts over time may reflect changes made by countries, the database provider, and/or administrative decisions. For example, it is expected that the wos_b_2020 database contains a higher number of publications for the most recent years than the wos_b_2019 database due to the continued indexing of items by Clarivate past the annual point in April at which data is cut to create the KB databases.

Increases in publications over time also result from both the continued growth of the national science systems and WoS' ongoing indexation over time, while sharp increases for a particular country may represent an actual increase in the number of a country's articles published in WoS-indexed journals, such as due to policy decisions, or reflect the recent indexing of a region- or country-specific journal. Decreases may reflect the de-indexation of a discipline-specific journal in which an entity commonly publishes or the stagnation of a sector, such as due to funding or policy decisions or the de-commissioning of an institution. Substantial deviations between databases – particularly in earlier years – or decreases in the current database in recent years may warrant investigation.

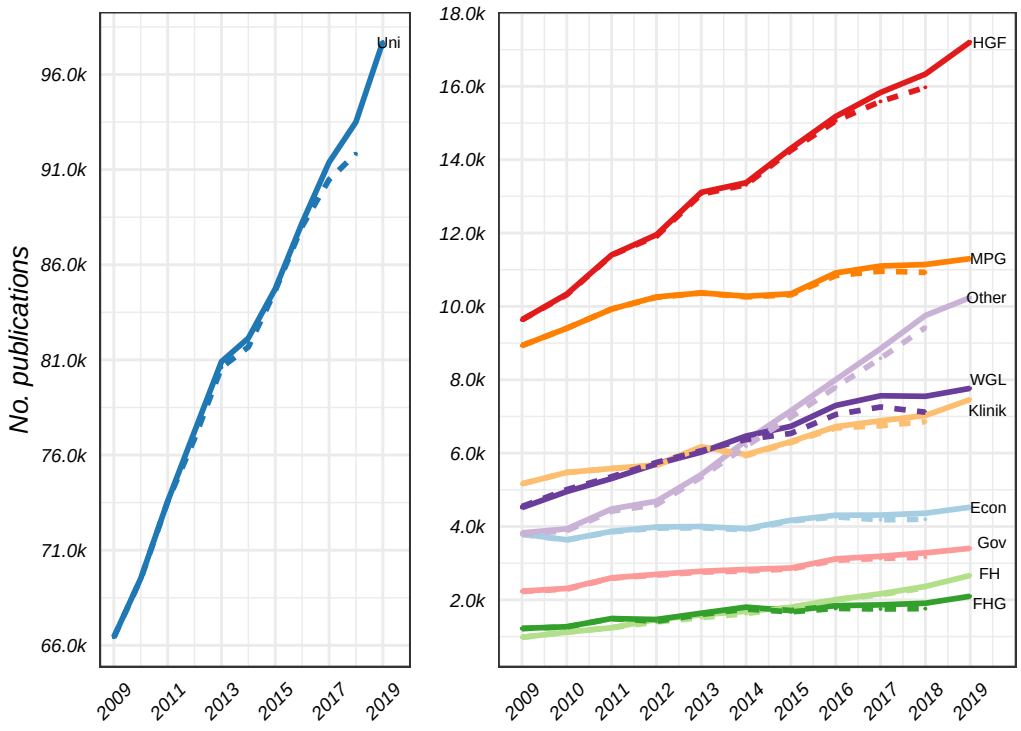

Figure 1: Whole counts of sectoral publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels' different scales.

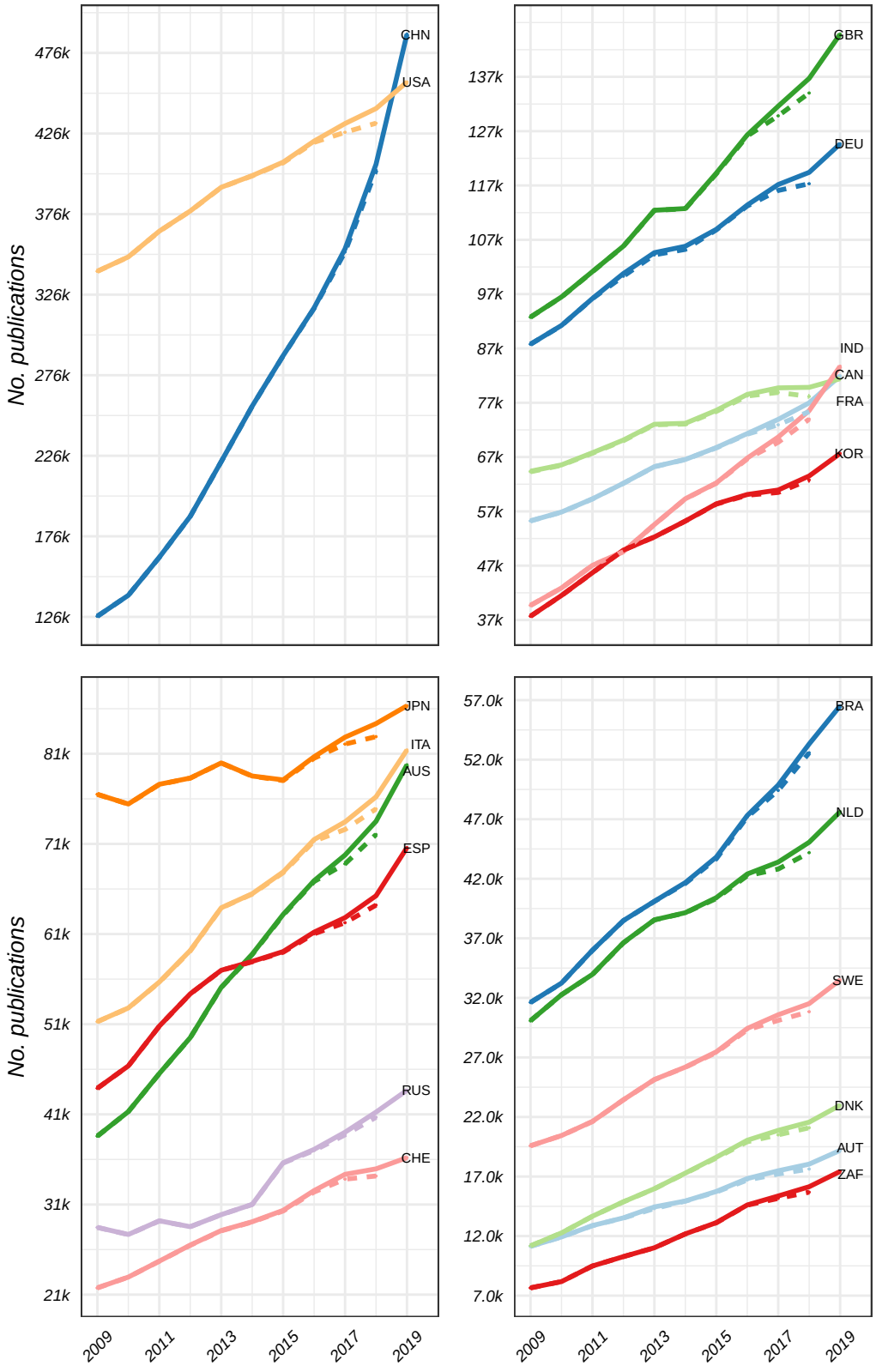

Figure 2: Whole counts of national publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels have different axes.

Excellence Rates: Selected countries and German sectors

Excellence Rates (ER) identify the percentage of an entity's publications that are in the 10% most highly cited publications from each discipline and could be considered of excellent quality on this basis. ERs are a common indicator used to assess an entity's performance, with an ER exceeding the expected 10% threshold interpreted as better than expected performance. ERs are calculated here based on the sc_traditional discipline classification. The ERs for the common years of the two databases up to 2016 are presented for German sectors in Figure 3 and for countries in Figure 4. As with whole counts of publications, we would expect general agreement between the databases, particularly in the earlier years of the time-series, so substantial deviations may warrant further analysis.

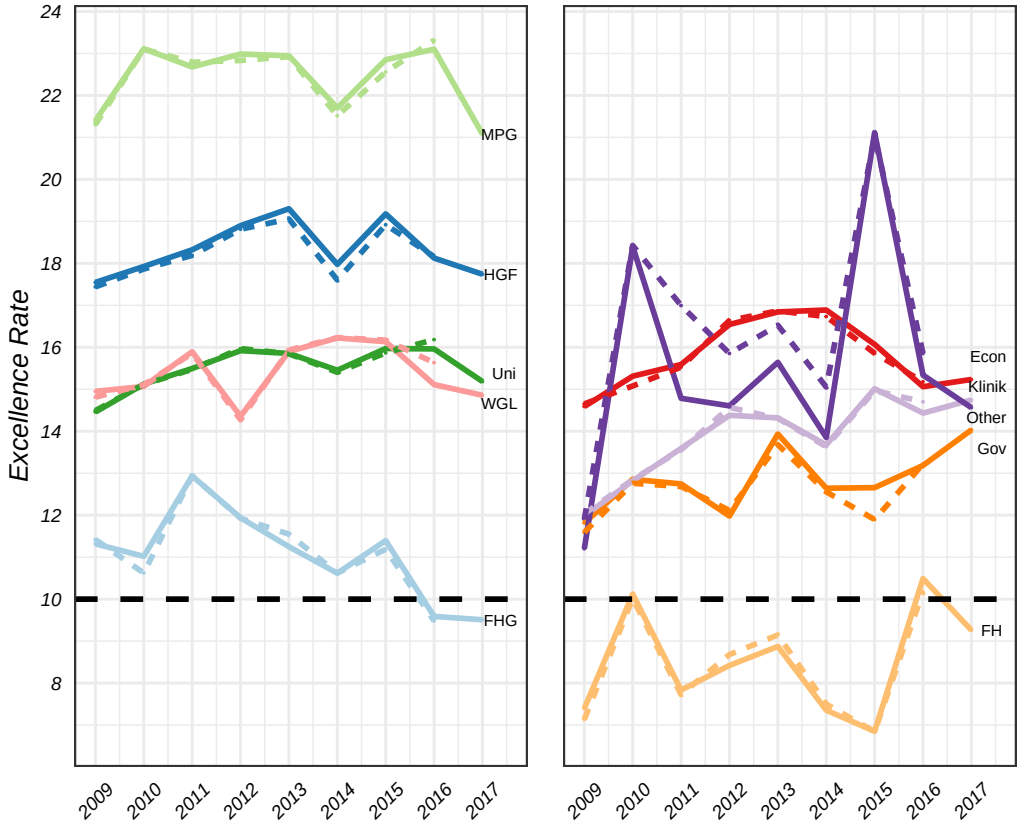

Figure 3: Excellence rates by sector, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

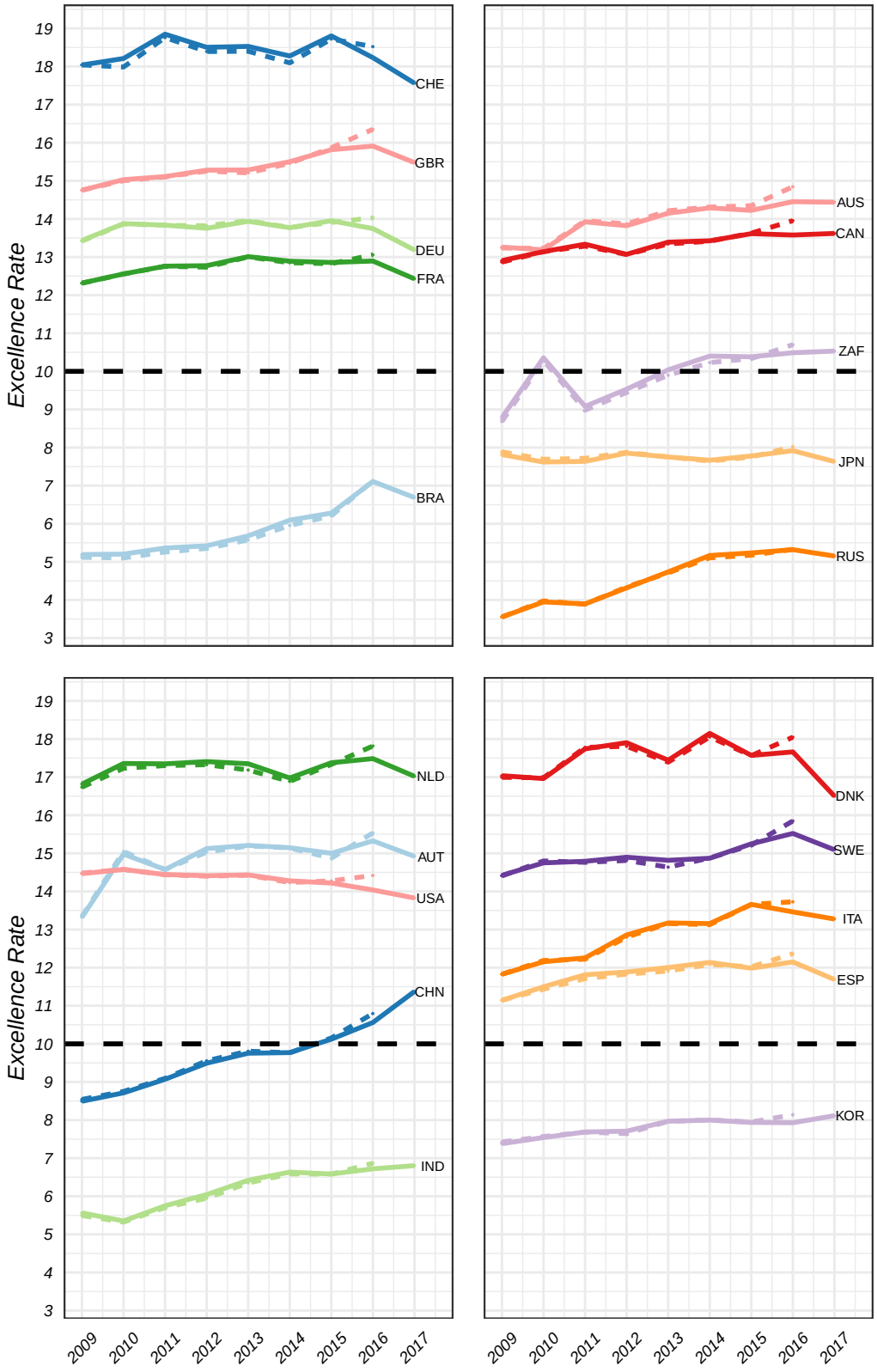

Figure 4: Excellence rates for selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

Excellence Rates: Thresholds by discipline

ERs are dependent on the number of citations a publication receives in relation to the threshold it must exceed to reach the top 10% of the pool of reference publications. A change in the 10% threshold for a discipline can make it more or less difficult for a publication to exceed the threshold, which can have knock-on effects for a sector or country's ER over time. For example, substantial differences in countries' ERs between WoS and Scopus were observed in Stahlschmidt, Stephen and Hinze [1] due to the differences in coverage between the two databases, as Scopus' greater coverage of more sparsely cited journals lowers the ER threshold and allows high-performing countries to receive higher ERs. The higher consistency of coverage in WoS, compared to between WoS and Scopus, means we expect less change in the ER thresholds between the iterations of the WoS databases, however changes in the journals indexed may influence the ER threshold for disciplines, potentially affecting the ERs of countries or, in particular, sectors due to their stronger disciplinary focus.

To examine changes in thresholds, we present in Figure 5 the ER thresholds for articles and reviews in each sc_traditional discipline. We assess articles and reviews separately given the known differences in citation patterns between the document types. Large increases in the threshold would require publications to achieve substantially more citations to exceed the 10% threshold and be included in the ER, while a decrease in the threshold means publications require fewer citations than previously.

In the top panels of Figure 5 we see the ER thresholds for each discipline in 2016 in both the wos_b_2019 and wos_b_2020 databases. The colour denotes the number of disciplines with each combination of thresholds, from fewer in blue to more in red. These panels depict the changes in ER thresholds in the same year between databases, providing context for any differences observed in 2016 in Figures 3 and 4. In the bottom panels we present again the thresholds for each discipline in 2016 in the wos_b_2019 database but now compared against the threshold in 2017 in the wos_b_2020 database. These panels highlight changes between the latest years in each database, indicating whether we could expect to see changes in ERs between the databases.

The outlying disciplines with the greatest change in thresholds in the bottom panels of Figure 5 are shown in Tables 1 and 2, along with disciplines where the previous threshold was zero, highlighting potentially new or emerging disciplines.

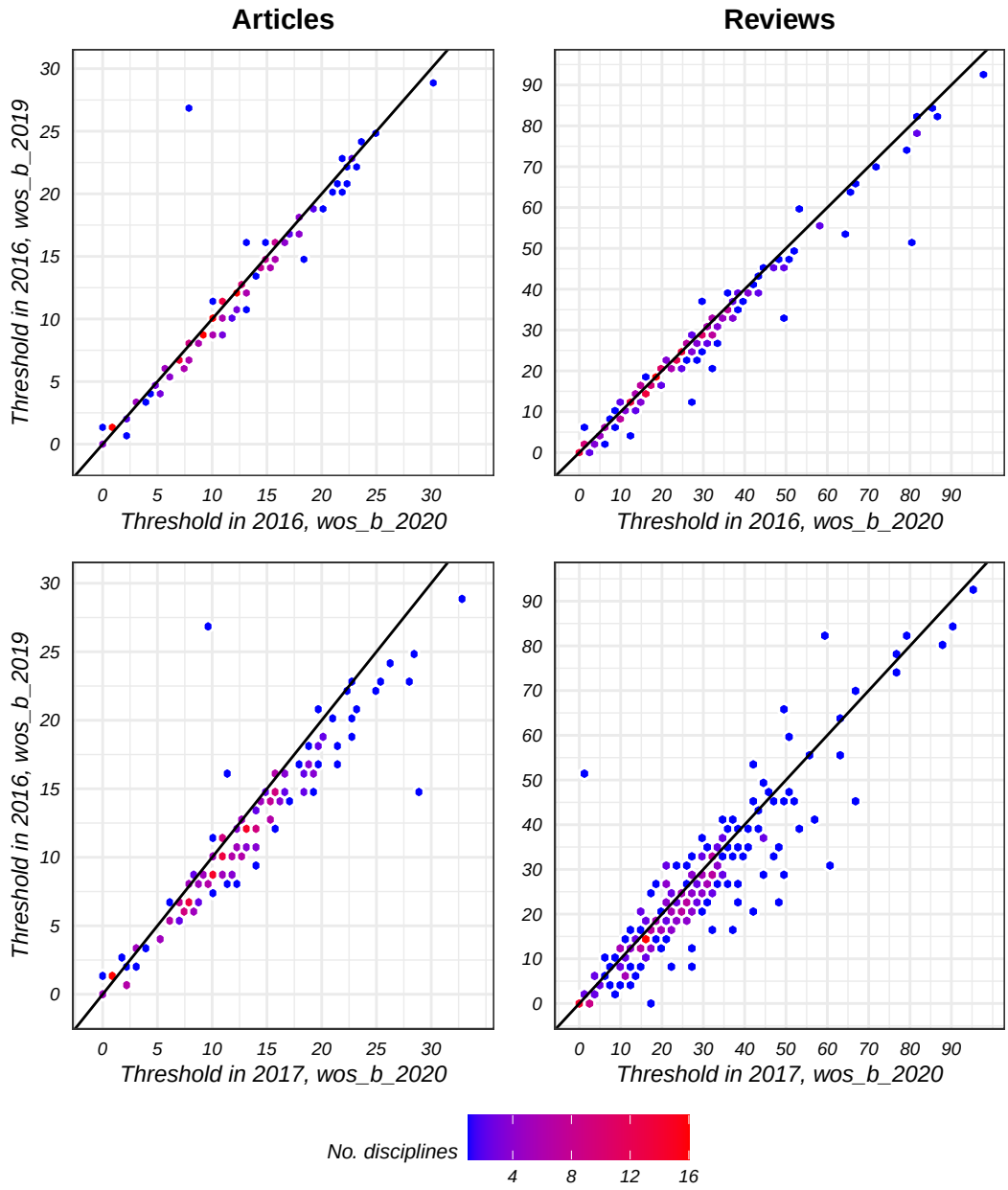

Figure 5: The ER threshold for articles and reviews in each discipline (sc_traditional) between databases, where colour denotes the number of disciplines with this combination of thresholds.

Table 1: Articles: Disciplines where the ER threshold decreased, or increased by over 40% between 2016 in wos_b_2019 and 2017 in wos_b_2020, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Art	1	2	2,845	100.0
Film, Radio, Television	1	2	996	100.0
History	1	2	7,384	100.0
Humanities, Multidisciplinary	1	2	3,883	100.0
Religion	1	2	3,971	100.0
Computer Science, Cybernetics	15	29	1,736	93.3
Regional & Urban Planning	9	14	1,370	55.6
Hospitality, Leisure, Sport & Tourism	8	12	2,984	50.0
Philosophy	2	3	6,108	50.0
Engineering, Aerospace	7	10	3,776	42.9
Astronomy & Astrophysics	21	20	20,222	-4.8
Crystallography	11	10	6,292	-9.1
Anatomy & Morphology	9	8	2,139	-11.1
Medicine, Legal	9	8	1,908	-11.1
Ornithology	7	6	1,195	-14.3
Physics, Nuclear	16	11	5,906	-31.2
Psychology, Psychoanalysis	3	2	494	-33.3
Planning & Development	27	10	1,709	-63.0
Poetry	1	0	152	-100.0

Table 2: Reviews: Disciplines with a current ER threshold of at least 10, where the threshold decreased by over 25%, increased by over 60% between 2016 in wos_b_2019 and 2017 in wos_b_2020, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
History Of Social Sciences	0	3	36	Inf
Literary Theory & Criticism	0	1	8	Inf
Literature, American	0	1	10	Inf
Literature, British Isles	0	2	9	Inf
Logic	1	17	3	1,600.0
Development Studies	8	28	61	250.0
Cultural Studies	3	10	22	233.3
Language & Linguistics	4	13	63	225.0
Social Issues	9	23	37	155.6
Microscopy	16	37	55	131.2
Regional & Urban Planning	13	28	46	115.4
Transportation Science & Technology	23	49	50	113.0
Computer Science, Theory & Methods	16	32	107	100.0
Engineering, Aerospace	21	41	68	95.2
Quantum Science & Technology	32	61	29	90.6
Linguistics	7	13	157	85.7
Women's Studies	6	11	41	83.3
Materials Science, Paper & Wood	22	38	60	72.7
Computer Science, Artificial Intelligence	29	49	186	69.0
Anthropology	6	10	116	66.7
Chemistry, Inorganic & Nuclear	66	49	478	-25.8
Andrology	15	11	177	-26.7
Gerontology	26	19	172	-26.9
Agricultural Engineering	82	59	105	-28.0
Statistics & Probability	20	14	72	-30.0
Ornithology	25	17	25	-32.0
Materials Science, Ceramics	31	21	76	-32.3
Physics, Fluids & Plasmas	31	21	88	-32.3

Citations: Mean 3-year citations of articles and reviews by discipline

The number of citations a publication could be expected to receive is dependent on its discipline. As such, we examine here the mean 3-year citations of articles and reviews by discipline. Mean 3-year citations (MC3) are the mean citations publications in each discipline accrue in the first 3 years after publication. As we did with ERs, we examine here in Figure 6 the last common year in both databases (top panels) to assess the retroactive effects stemming from changes made in the latest database, and the latest complete year in both databases (bottom panels) to assess potential structural changes and updates to the time-series. A greater deviation of disciplines from the central line indicates a greater degree of change in the mean citations of a discipline's items between years. Data are based on the sc_traditional discipline classification. The outlying disciplines from the bottom panels of Figure 6 are shown in Tables 3 and 4, along with disciplines where the previous threshold was zero. We use a threshold of a current MC3 of at least 1 for articles and 3 for reviews to remove disciplines with spurious changes due to low level of citations.

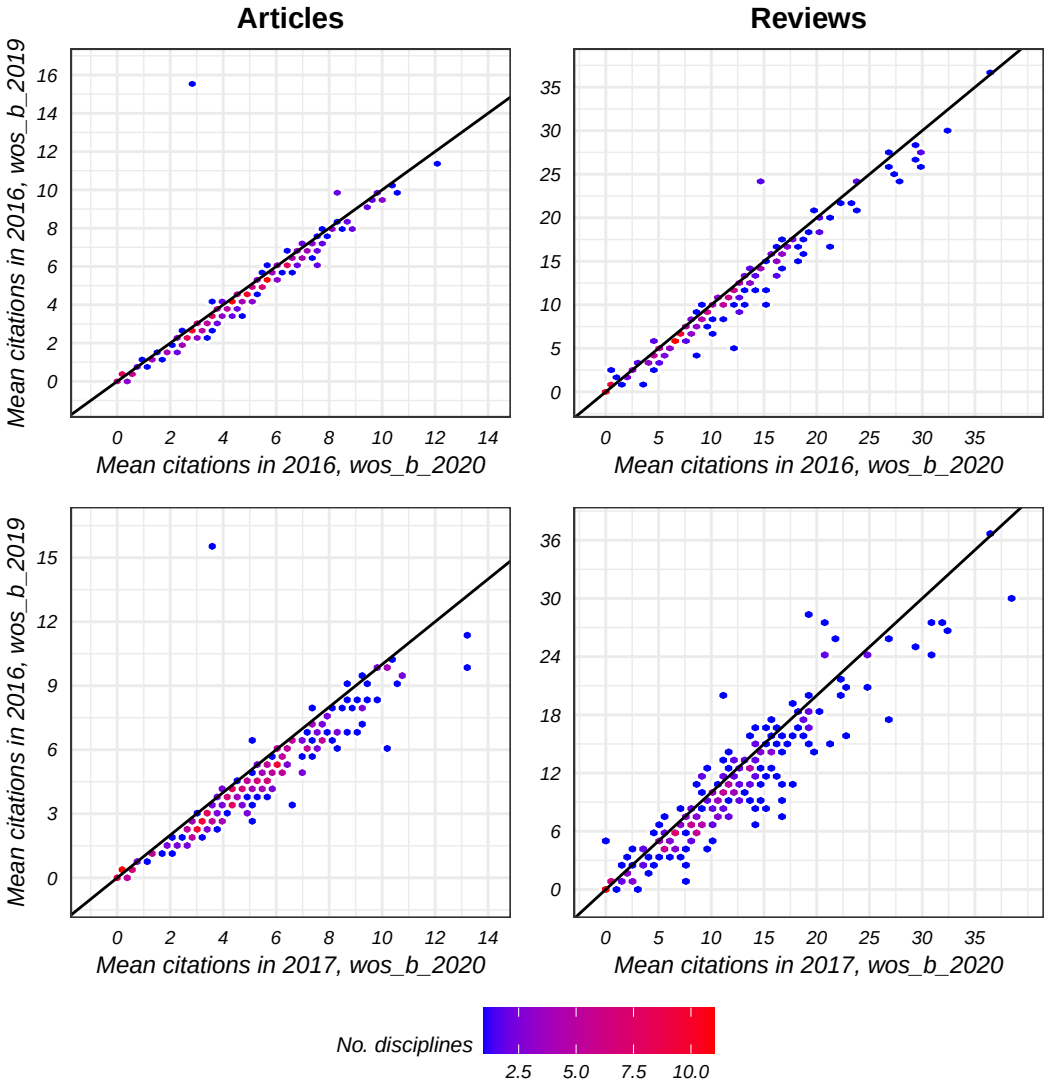

Figure 6: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

Table 3: Articles: Disciplines with a current MC3 of at least 1, where the MC3 decreased by over 20% or increased by over 50% between 2016 in wos_b_2019 and 2017 in wos_b_2020, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Regional & Urban Planning	3.3	6.5	1,370	98.2
Hospitality, Leisure, Sport & Tourism	2.7	5.0	2,984	84.2
Education & Educational Research	1.5	2.6	11,804	71.9
Computer Science, Cybernetics	6.2	10.2	1,736	64.5
Industrial Relations & Labor	1.8	3.0	1,053	64.0
Urban Studies	3.1	4.9	2,448	62.1
Language & Linguistics	0.7	1.1	4,507	61.2
Law	1.2	2.0	4,436	61.2
Engineering, Marine	2.3	3.8	1,357	60.7
Social Sciences, Interdisciplinary	1.7	2.7	5,833	60.7
Communication	2.2	3.4	3,913	58.3
Geography	3.1	4.9	4,541	57.9
Engineering, Multidisciplinary	3.3	5.1	12,529	53.5
Demography	2.1	3.2	1,141	53.4
Physics, Nuclear	6.6	5.1	5,906	-22.8
Planning & Development	15.5	3.6	1,709	-76.6

Table 4: Reviews: Disciplines with a current MC3 of at least 3, where the MC3 decreased by over 20% or increased by over 60% between 2016 in wos_b_2019 and 2017 in wos_b_2020, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Literary Theory & Criticism	0.0	0.1	8	Inf
Literature, American	0.0	0.2	10	Inf
Literature, British Isles	0.0	0.3	9	Inf
Literature, Slavic	0.0	0.1	14	Inf
Logic	1.0	7.6	3	658.3
Development Studies	2.4	8.0	61	234.2
Women's Studies	1.6	4.3	41	160.1
Regional & Urban Planning	4.1	9.4	46	127.3
Medical Ethics	3.0	6.7	24	122.5
Acoustics	7.6	16.5	100	116.4
Engineering, Multidisciplinary	6.8	13.7	254	100.9
Automation & Control Systems	8.7	16.8	98	93.5
Computer Science, Cybernetics	5.2	9.7	21	88.2
Materials Science, Paper & Wood	4.7	8.5	60	81.6
Anthropology	2.2	3.9	116	81.3
Linguistics	2.7	4.7	157	76.4
Engineering, Aerospace	8.8	15.3	68	73.0
Criminology & Penology	3.4	5.9	153	72.8
Social Issues	4.6	7.8	37	70.0
Mathematics, Applied	3.8	6.3	46	64.5
Transportation Science & Technology	8.5	13.9	50	63.1
Social Work	3.1	5.0	149	62.8
Public Administration	6.3	5.0	41	-20.4
Parasitology	14.5	11.5	429	-20.6
Gerontology	11.8	9.1	172	-22.9
Chemistry, Inorganic & Nuclear	27.4	20.6	478	-25.0
Mineralogy	12.7	9.4	126	-26.5
Andrology	7.5	5.4	177	-28.1
Imaging Science & Photographic Technology	28.1	19.6	36	-30.2
Physics, Fluids & Plasmas	20.4	11.2	88	-45.1

Uncited articles and reviews: Percent by selected countries and German sectors

While ERs represent the most highly cited publications and mean citations tell us about what's average, the percentage of uncited publications can tell us about the entities at the tail end of the citation distribution. When examining uncited publications, we expect to see a decreasing trend in uncited publications over time. This occurs because citation counts are based on the items indexed in each database and so, as Clarivate continues to index journals, it increases the likelihood that any publication will have been cited by the indexed items. In particular, we would expect that the percentage of uncited publications in the last common year would be lower in the current database than the previous database, as data added in the latest iteration "complete" the incomplete last year of the previous database. An increase in uncited publications in the latest year may reflect processing issues that require investigation. We present in Figures 7 and 8 the percentage of articles and reviews per German sector and selected country that remained uncited 3 years after they were published.

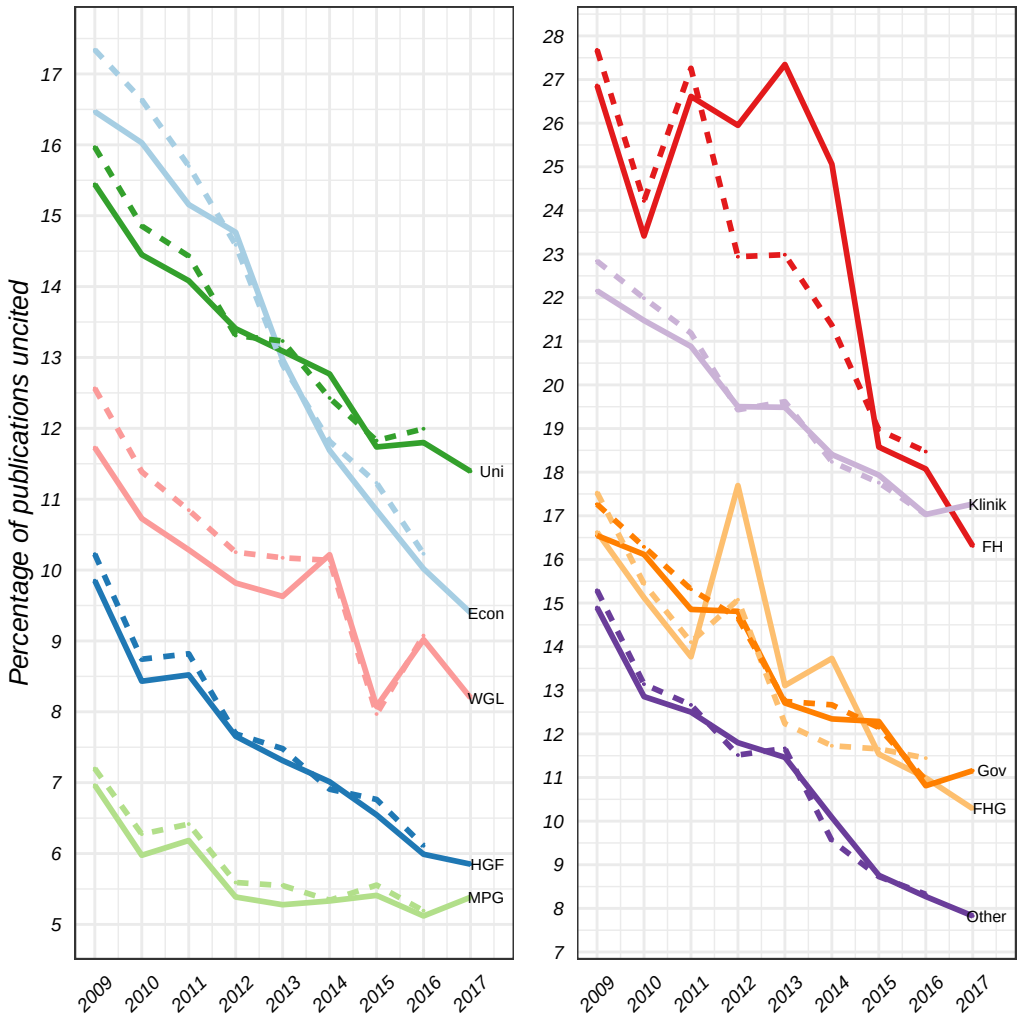

Figure 7: The percentage of uncited publications by German sector, based on whole counts, where dashed lines show the previous database and full lines show the current database.

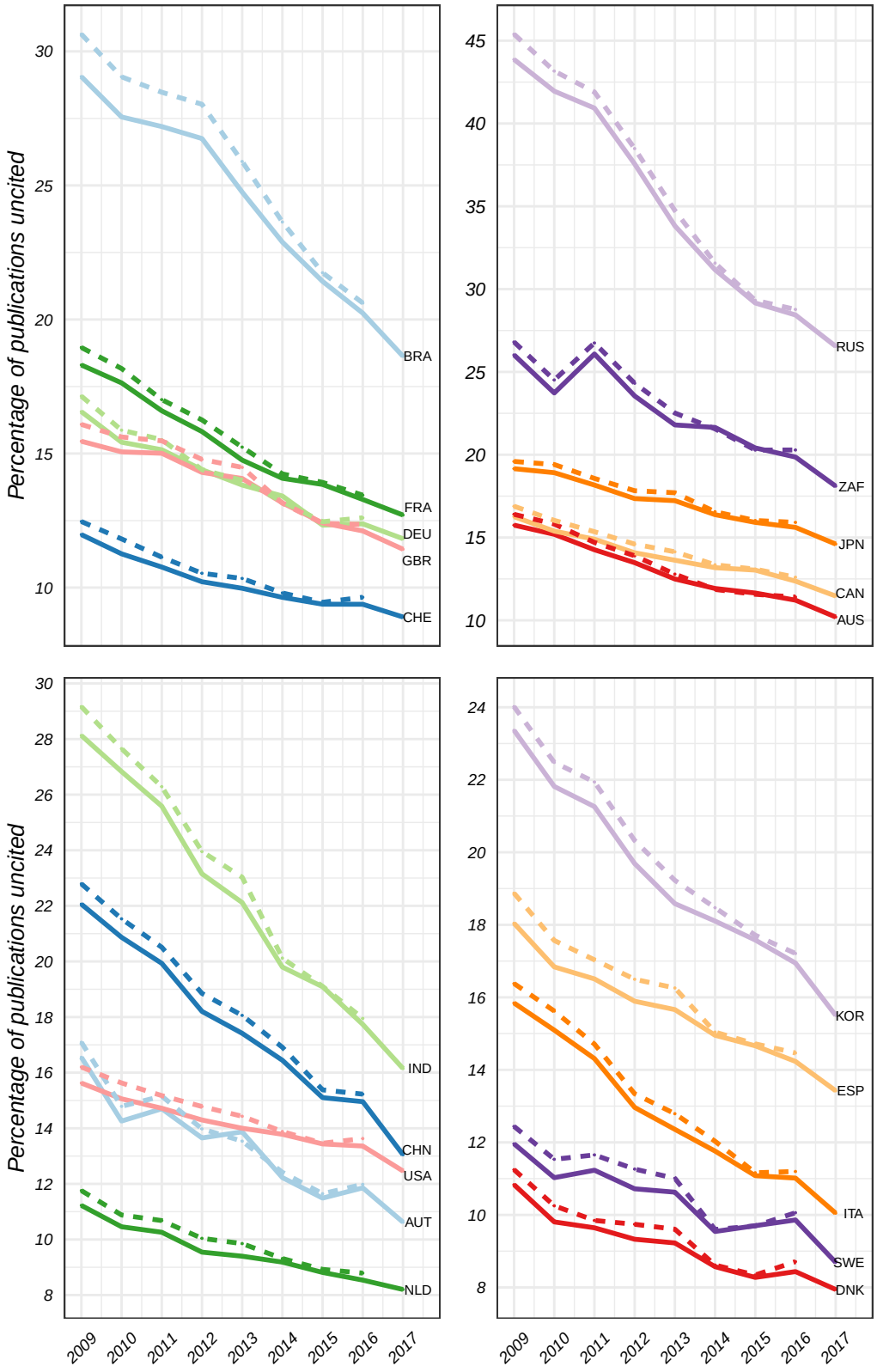

Figure 8: The percentage of uncited publications by selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database.

Disciplines: Changes in discipline classification

The two tables in this section highlight whether any changes have been made to WoS' discipline classifications, the sc_traditional in Table 5 and sc_extended in Table 6. This could include splits, aggregations or removals of a discipline, or the inclusion of a new discipline to reflect new and emerging topics. We identify changes in the classification structure by comparing the number of articles and reviews attributed to each discipline in the latest years of each database and selecting those disciplines where the number was zero in one year but not in the other.

Disciplines with no prior publications but some in the current year suggest the discipline may have been recently added, while the opposite suggests the discipline may have been removed or merged. Changes may also reflect changes in spelling or punctuation of the discipline name. Any changes should be checked with WoS' published classification structure.

Table 5: Changes in the sc_traditional discipline classification structure between the previous and current databases.

Classification	Previous pubs	Current pubs

Table 6: Changes in the sc_extended discipline classification structure between the previous and current databases.

Classification	Previous pubs	Current pubs

Disciplines: Changes in articles and reviews by discipline

This section identifies the disciplines that had a substantial change in the number of publications assigned to them between the latest years in each database. Changes in counts of publications per discipline reflect changes in the journals indexed, the classification structure, and any potential processing issues. As such, any large changes shown here may be worth examining.

We show in Figure 9 the 20 disciplines with the highest percentage increases and decreases in publication counts between 2018 in wos_b_2019 and 2019 in wos_b_2020. The number shown next to each bar is the numerical change in publication counts. We have used whole counting and the disciplines are based on the sc_traditional classification. Disciplines previously identified as being new or removed have not been included here.

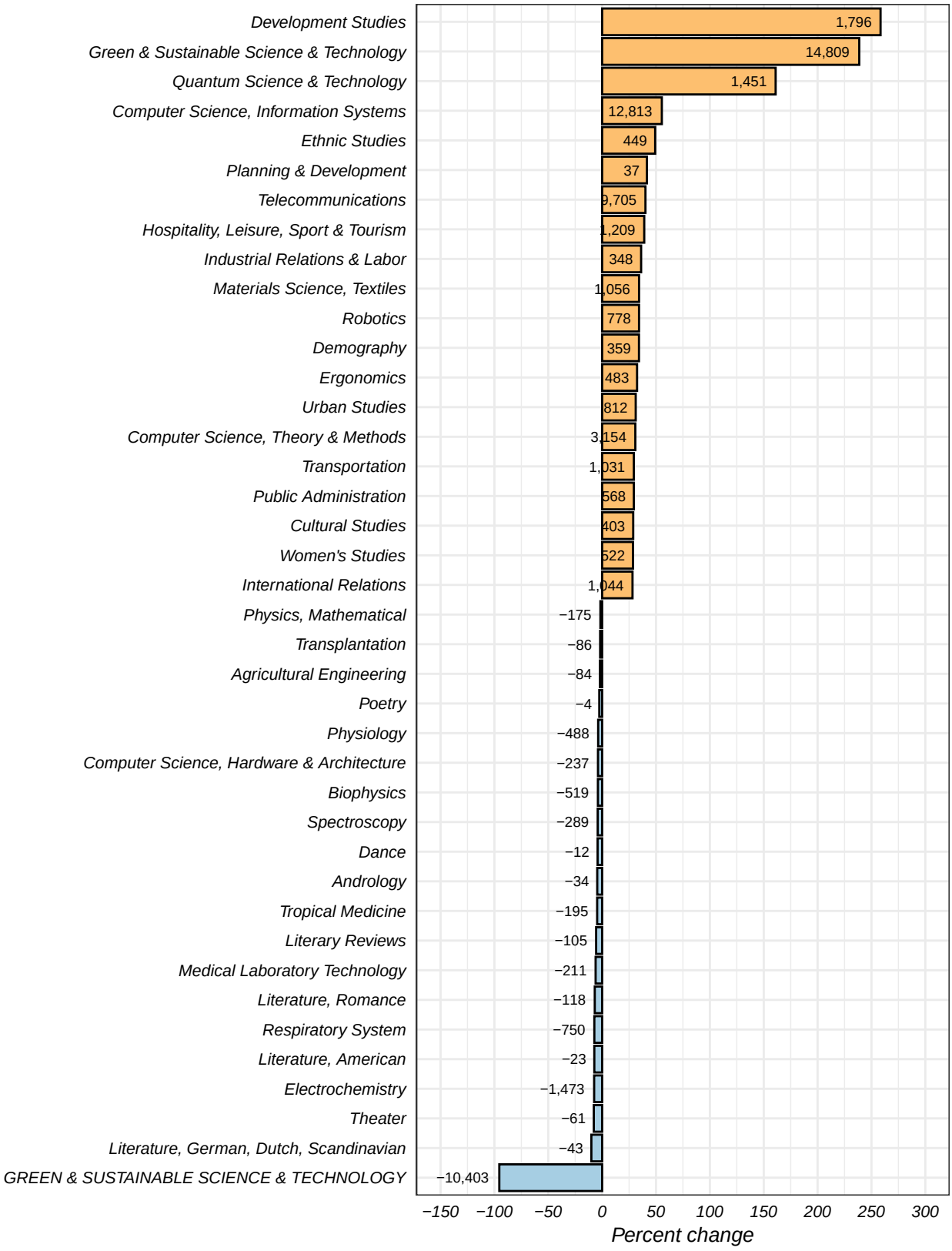

Figure 9: The 40 disciplines with the highest percentage change in publication counts between 2018 in wos_b_2019 and 2019 in wos_b_2020, with numerical difference in counts.

Disciplines: Number of publications not assigned to a discipline

This section presents in Figure 10 the percentage of publications in each database that were not assigned to a discipline over the previous 10 years. Complete assignment of publications to disciplines is important as citation-based indicators typically use field-normalisation to account for differences in citation practices between disciplines. As such, items missing discipline information are excluded from such analyses and so large percentages of, or large changes in, unclassified items should be investigated.

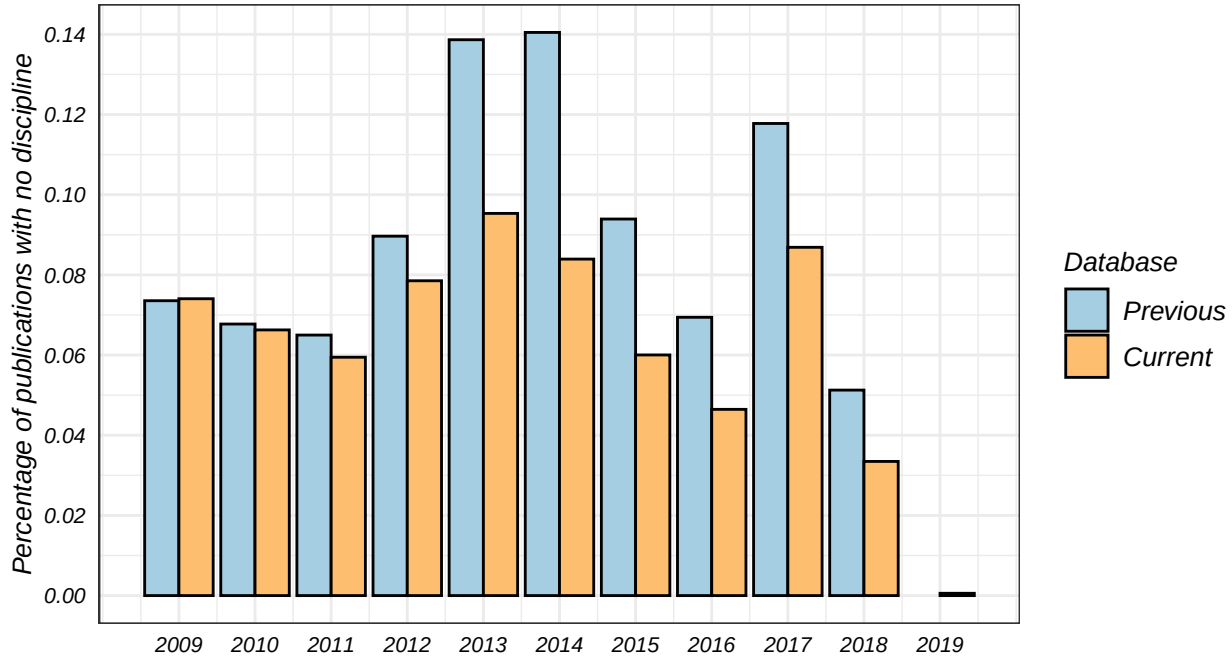

Figure 10: The percentage of publications in each database that do not have a discipline classification.

Metadata: Changes in pubyear, doctype and pubtype

This section details the number of items for which changes were made to key metadata in the latest iteration of the database. We look at changes in the recorded publication year, document type and publication type as these three variables are typically the key inclusion criteria for bibliometric analyses. A change in metadata for a large number of items may be problematic, particularly if the changes are not randomly distributed, such as adjustments having been made to items from a particular journal or set of publications, which may affect counts and indicators for specific entities. Some changes can be expected as Clarivate updates or corrects items, however a large number of items or a change in a time-series may require investigation.

We identify changes in the metadata of in-scope items by first matching items between the wos_b_2019 and wos_b_2020 databases using the UT_EID identifier and then counting the number of instances where matched items do not have the same publication year, document type (i.e. an article or review has been changed to a different document type) or publication type (i.e. the publication type changed from journal to another type) between databases. As such, Table 7 shows the number of items that have had their metadata changed between the previous and current databases. Data are presented based on the publication year recorded in the previous database.

Table 7: The number of items with changes in metadata between the previous and current database versions.

Year	Pub. year	Doc. type	Pub. type
2009	0	268	0
2010	0	76	0
2011	5	40	0
2012	0	46	0
2013	2	52	0
2014	1	47	0
2015	31	63	0
2016	18	213	0
2017	34	568	0
2018	857	522	0

Metadata: Publications from incorrect indices

The KB contract with Clarivate specifies that we receive data from the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (AHCI). The inclusion of items from other indices, such as the Emerging Sources Citation Index, can be problematic as these items may fundamentally differ from those in the three core indices in, for instance, the countries of their authors and publishing journals, which can influence citation-based indicators. As such, we include a check here whether the database includes items indexed outside of the SSCI, SCIE and AHCI. Table 9 shows the annual number of items from non-core idices present in the wos_b_2019 database, and Table **??** shows the same for the wos_b_2020 database. As such, blank tables indicate there are no incorrect inclusions.

Table 8: Number of publications indexed in out of scope indices, wos_b_2019.

PUBYEAR	WOS.BSCI	WOS.ESCI
2015	8	NA
2009	NA	299
2010	NA	384
2011	NA	263
2012	NA	713
2013	NA	278
2014	NA	202
2015	NA	16372
2016	NA	88244
2017	NA	55716

Table 9: Number of publications indexed in out of scope indices, wos_b_2020.

PUBYEAR

Metadata: Missing metadata variables

The section presents the annual percentage of publications in each database that are missing particular metadata, including page numbers, journal issue and volume information, DOIs, titles, references, abstracts, and keywords. Data for wos_b_2019 are in Table 10 and wos_b_2020 are in Table 11. We could reasonably expect improvements over time in missing metadata, such as for DOIs through increasing uptake of this identifier, however increasing missing metadata should be investigated. NAs indicate there were no items missing this metadata, while zeroes indicate there were some items with metadata missing, but less than 0.1%.

Table 10: Percentage of publications missing metadata, wos_b_2019.

Year	No page	No issue	No vol.	No DOI	No title	No refs	No abs.	No keys
2009	NA	5.0	2.0	20.0	NA	1.3	NA	NA
2010	NA	5.2	1.8	17.5	NA	1.2	NA	0
2011	NA	5.7	1.9	15.3	NA	1.1	NA	NA
2012	NA	11.6	2.3	13.0	NA	1.1	NA	0
2013	NA	16.0	2.4	10.5	NA	1.1	NA	0
2014	NA	20.2	2.5	9.6	NA	0.9	NA	NA
2015	NA	22.5	2.2	8.7	NA	0.8	NA	NA
2016	NA	25.0	1.9	6.9	NA	0.7	NA	NA
2017	NA	26.4	1.8	5.5	NA	0.7	NA	0
2018	NA	27.9	1.9	4.1	NA	0.5	NA	NA

Table 11: Percentage of publications missing metadata, wos_b_2020.

Year	No page	No issue	No vol.	No DOI	No title	No refs	No abs.	No keys
2009	NA	5.0	2.0	20.0	NA	1.3	NA	0
2010	NA	5.2	1.7	17.5	NA	1.2	NA	0
2011	NA	5.7	1.8	15.3	NA	1.1	NA	NA
2012	NA	11.6	2.2	12.9	NA	1.1	NA	0
2013	NA	16.0	2.3	10.5	NA	1.1	NA	0
2014	NA	20.2	2.4	9.6	NA	0.9	NA	NA
2015	NA	22.5	2.0	8.7	NA	0.8	NA	NA
2016	NA	25.0	1.7	6.9	NA	0.7	NA	NA
2017	NA	26.4	1.6	5.5	NA	0.7	NA	0
2018	NA	27.8	1.7	4.2	NA	0.5	NA	0
2019	NA	30.4	4.5	3.2	NA	0.4	NA	0

Institution and country data: Number of articles and reviews with missing data

Bibliometric analyses often examine indicators at the level of institutions or countries. Further, fractional counting can be applied based on institutions, with articles apportioned according to authors' affiliations. As such, it is imperative for accurate indicators that most, if not all, items have institution and country data, as missing information removes otherwise valid items from analyses.

The Items table of the bibliometric databases holds a record of all available items, while the associated data about authors' affiliations are held, in part, in the Institutions table. We have operationalised missing institution information here as publications that appear in the Items table but have no corresponding information in the Institutions table. We present in the top panel of Figure 11 the number of items in each database between 2009 and 2018 with no institution information. Additionally, items can have institution information but no country code – from which country counts are derived – and these are shown in the bottom panel of Figure 11. Large disparities between the databases or substantial increases in missing information should be investigated.

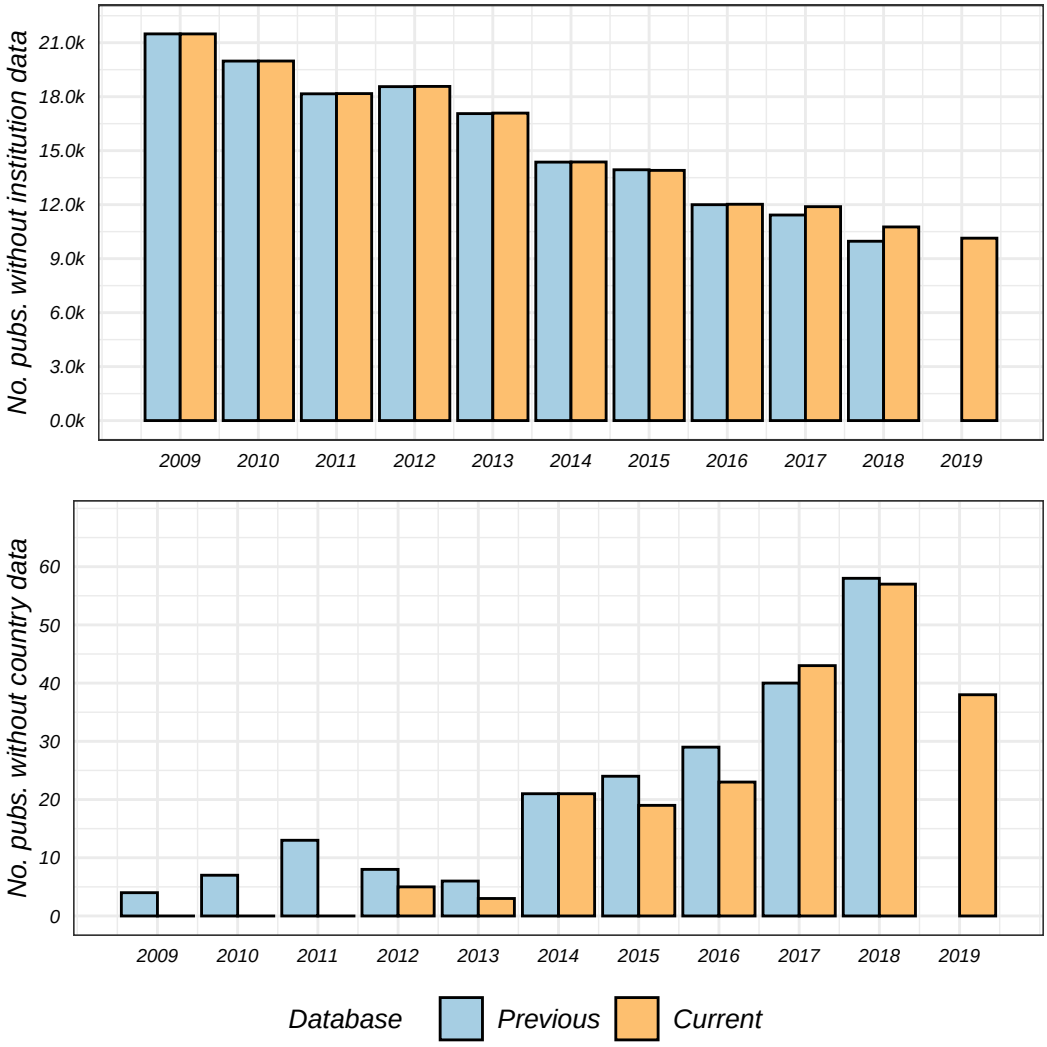

Figure 11: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

Author-institution links: Percentage complete by Research Area and discipline

Similarly to ensuring that all or most items have institution and country information, it is important for allocating publications to entities that authors' affiliations with institutions have been assigned for the majority, or ideally all, items. As such, we examine here the percentage of items in each sc_extended discipline with complete links between authors and institutions.

In Figure 12, we see in the left panel the percentage of complete links for 2018 data in both the previous and current databases, highlighting any retroactive changes that may have been made in the current database. In the right panel is again the percentage of complete links made in 2018 in the wos_b_2019, now compared with the 2019 in the wos_b_2020, indicating potential changes between the latest year in each database.

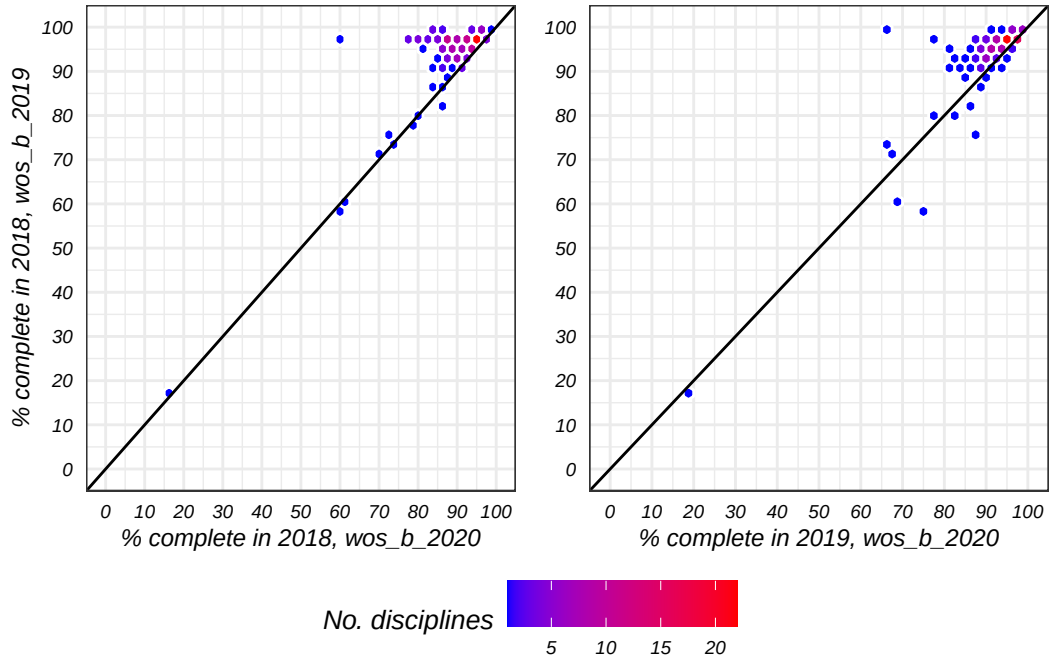

Figure 12: The percentage of complete author-institution links by disciplines (sc_extended).

The outlying disciplines observed in the right panel of Figure 12 that have a change of more than 7 percentage points in the percentage of complete author-institution links between databases are shown in Table 12.

Table 12: Disciplines (sc_extended) with a change of more than 7 percentage points in missing links between 2018 in wos_b_2019 and 2019 in wos_b_2020.

Discipline	Prvs items	% prvs complete	Crrnt items	% crrnt complete	Change
Life Sciences & Biomedicine - Other Topics	1,864,580	98.3	2,909,116	65.7	-32.6
Evolutionary Biology	170,097	97.5	174,478	77.4	-20.0
Rehabilitation	140,542	95.3	338,468	80.9	-14.4
Materials Science	10,449,300	98.0	15,199,281	87.2	-10.7
Allergy	73,038	93.8	82,156	83.5	-10.3
Criminology & Penology	25,457	96.8	82,127	87.3	-9.5
Nursing	189,658	90.0	203,577	80.7	-9.3
Biomedical Social Sciences	69,745	95.1	92,317	86.0	-9.1
Chemistry	14,964,872	97.9	16,099,579	89.0	-8.8
Hematology	228,105	93.9	258,809	85.4	-8.5
Crystallography	307,830	98.0	341,724	90.1	-7.9
Development Studies	9,700	96.7	36,070	88.9	-7.8
Cell Biology	2,894,600	98.4	3,860,177	91.3	-7.1
Architecture	36,785	61.1	36,164	69.8	8.7
Art	33,713	75.1	61,942	86.5	11.4
Film, Radio & Television	7,879	58.1	11,101	74.1	15.9

To provide context to the percentage of complete links observed in the most recent years, in Figure 13 we present the percentage of complete links made between authors and affiliations in each Research Area over the last decade in both databases, plus 2019 in wos_b_2020. Substantial changes between years or differences between the databases may require investigation of the cause.

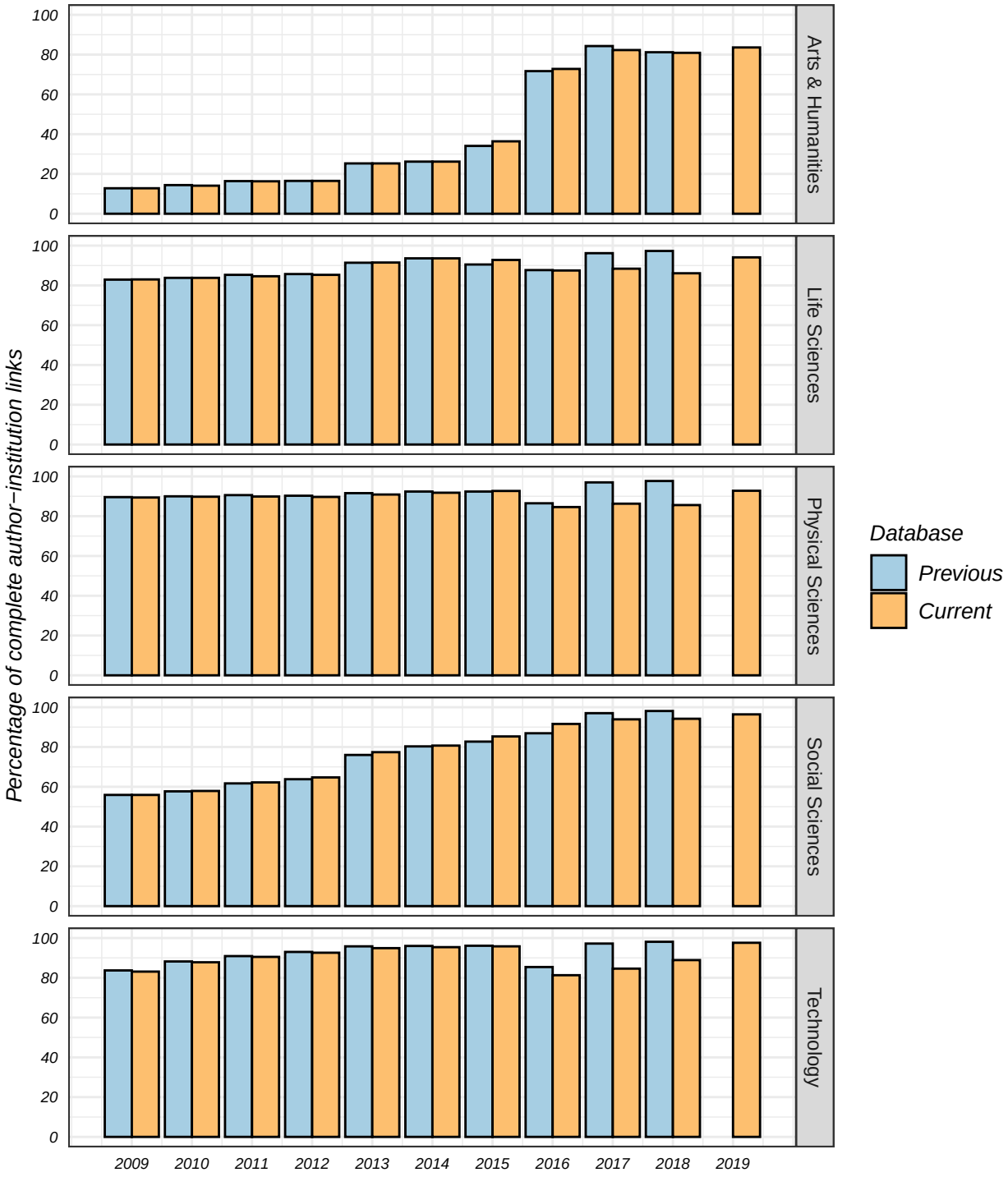

Figure 13: The annual percentage of complete author-institution links by Research Area and database.

German institutions: German publications missing from KB institution coding

In Figure 14 we show the annual number of German publications, i.e. those with a 'DEU' country code, that were not assigned a KB institution code through the I-Kodierung process. Increases over time are likely due to increasing publication rates and the foundation of new institutions that have not yet been integrated into the coding process. However, publications without KB institutions are typically excluded from sector-level analyses, so it is important to understand the extent of missing institution information.

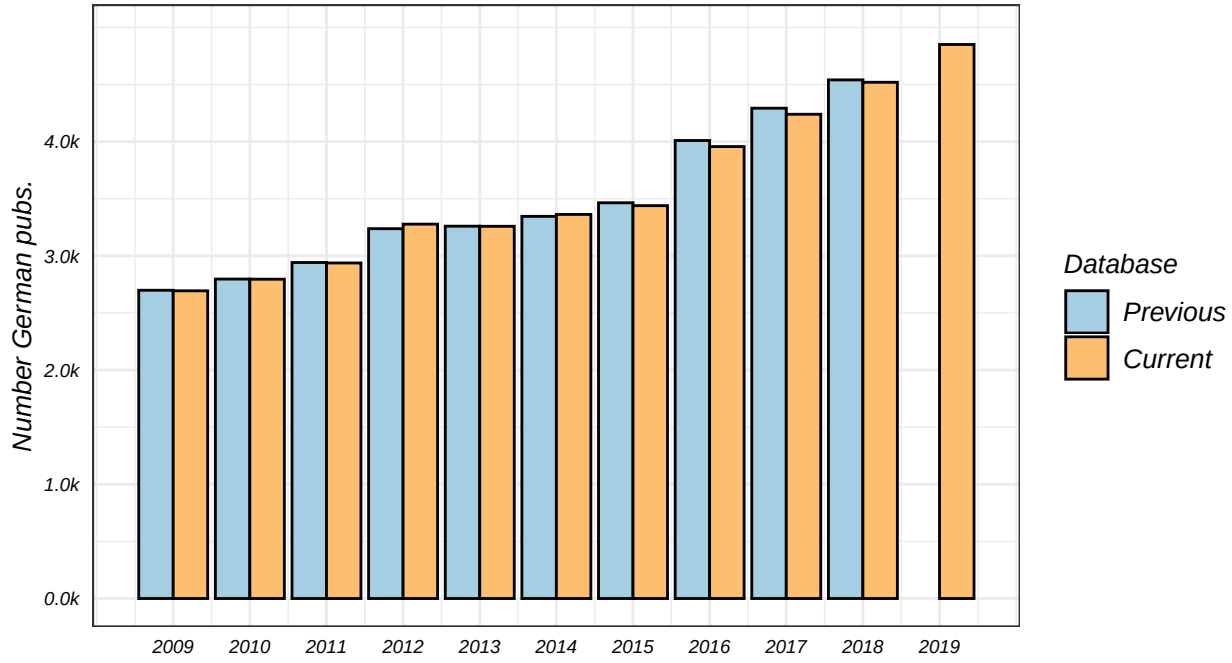

Figure 14: The number of German publications in each database that are missing a KB institution.

German institutions: Changes in whole counts of articles and reviews

This section compares changes in the number of articles and reviews published by German institutions between the latest years available in each database. These tables can assist in identifying institutions for which substantial numbers of publications have been added, removed or otherwise changed in the latest database. They can also aid in assessing the degree of change in publication numbers for larger institutions, which may require further examination if considered unusual or excessive.

Table 13 presents potentially new institutions – these had no publications in 2018 in the wos_b_2019 database but more than five publications in 2019 in the wos_b_2020 database. Conversely, Table 14 shows the institutions that had at least five publications in 2018 in the wos_b_2019 database but no publications recorded in 2019 in the wos_b_2020 database. We also highlight in Tables 15 and 16 the larger institutions (with at least 20 publications) that had a change in publication counts of more than 40% between 2018 and 2019 in the wos_b_2019 and wos_b_2020 databases.

Table 13: Institutions with more than 5 publications in 2019 in wos_b_2020 that had no publications in 2018 in the wos_b_2019 database.

PK_KB_INST	Name	Previous pubs	Current pubs
5,477	Leibniz-Institut fur Photonische Technologien e.V. (IPHT)	0	189
5,471	DWI - Leibniz-Institut fur Interaktive Materialien	0	112
5,431	Nationales Centrum fur Tumorerkrankungen / University Cancer Center	0	77
2,121	Fraunhofer-Institut fur Mikrostruktur von Werkstoffen und Systemen	0	67
1,328	Nokia Siemens Networks GmbH & Co. KG	0	39
5,478	Leibniz-Institut fur Werkstofforientierte Technologien - IWT	0	34
5,432	Centogene AG	0	33
5,483	Fraunhofer-Institut fur Energiewirtschaft und Energiesystemtechnik	0	25
835	Bundesinstitut fur Bevolkerungsforschung	0	23
29	Leibniz-Institut fur Raumbezogene Sozialforschung (IRS)	0	17
5,373	LungenClinic Grosshansdorf - Akademisches Lehrkrankenhaus der Universitat zu Lubeck	0	17
5,479	Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)	0	17
1,971	Technische Hochschule Koln	0	16
5,482	Leibniz-Zentrum Moderner Orient (ZMO)	0	14
5,488	Helmholtz-Institut fur Funktionelle Marine Biodiversitat an der Universitat Oldenburg	0	12
5,484	Fraunhofer-Einrichtung fur Additive Produktionstechnologien IAPT	0	9
5,430	IKDT - INSTITUT KARDIALE DIAGNOSTIK und THERAPIE GMBH	0	7
500	Bezirkskrankenhaus Augsburg Klinik fur Psychiatrie, Psychotherapie und Psychosomatik	0	6
5,436	ICH Study Center	0	6
5,437	Infectious Diseases Center Hamburg	0	6
5,476	Leibniz-Institut fur Ost- und Sudosteuropaforschung	0	6

Table 14: Institutions with no publications in 2019 in wos_b_2020 that had more than 5 publications in 2018 in the wos_b_2019 database.

PK_KB_INST	Name	Previous pubs	Current pubs
1,351	microfluidic ChipShop GmbH	6	0
4,172	Bernstein Fokus: Neurotechnologie (BFNT)	8	0
5,277	Hochschule fur angewandte Wissenschaften Wurzburg Schweinfurt	9	0

Table 15: Institutions with more than 20 publications in 2018 in the wos_b_2019 that increased in publication counts by over 40% to 2019 in the wos_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
35	Leibniz-Institut fur Oberflachenmodifizierung (IOM)	25	82	57	228.0
1,155	Fraunhofer-Institut fur Integrierte Systeme und Bauelementetechnologie	21	53	32	152.4
646	Hochschule fur angewandte Wissenschaften Coburg	21	41	20	95.2
492	Klinikum Bremen-Mitte gGmbH	29	55	26	89.7
15	Paul-Drude-Institut fur Festkorperelektronik (PDI)	42	79	37	88.1
604	Hochschule fur Wirtschaft und Recht Berlin	21	38	17	81.0
1,140	Fraunhofer-Institut fur Silicatforschung (ISC)	45	81	36	80.0
1,156	Fraunhofer-Institut fur Integrierte Schaltungen (IIS)	39	70	31	79.5
1,151	Fraunhofer-Institut fur Molekularbiologie und Angewandte Okologie	114	200	86	75.4
460	St.-Johannes-Hospital Dortmund	27	47	20	74.1
4,712	Institut fur Energie- und Umwelttechnik e.V.	22	38	16	72.7
5,290	Deutsches Zentrum fur Infektionsforschung	313	534	221	70.6
4,696	European XFEL GmbH	70	118	48	68.6
462	Klinikum Dortmund gGmbH	31	51	20	64.5
523	Klinikum Bayreuth GmbH	28	46	18	64.3
521	DRK Kliniken Berlin	31	50	19	61.3
804	Institut fur Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur fur Arbeit (BA)	52	83	31	59.6

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
56	Fraunhofer-Institut fur Optronik, Systemtechnik und Bildauswertung IOSB	22	35	13	59.1
1,152	Fraunhofer-Institut fur Lasertechnik (ILT)	34	54	20	58.8
4,428	Restkategorie Universitat en, Kunst- und MusikhochschulenHochschulen	44	69	25	56.8
648	Hochschule Bonn-Rhein-Sieg, University of Applied Sciences	36	55	19	52.8
340	Markus-Krankenhaus Akademisches Lehrkrankenhaus der Johann Wolfgang Goethe-Universitat Frankfurt am Main	45	68	23	51.1
654	Fachhochschule Bielefeld	47	71	24	51.1
152	Universitat Erfurt	49	74	25	51.0
1,069	Max-Planck-Institut fur Evolutionsbiologie	73	110	37	50.7
1,021	Max-Planck-Institut fur Stoffwechselforschung	44	66	22	50.0
1,103	GSI Helmholtzzentrum fur Schwerionenforschung	475	704	229	48.2
246	European Organisation for the Exploitation of Meteorological Satellites - EUMETSAT	27	40	13	48.1
30	Leibniz-Institut fur die Padagogik der Naturwissenschaften und Mathematik (IPN)	47	69	22	46.8
971	Forstliche Versuchs- und Forschungsanstalt Baden-Wurttemberg (FVA)	30	44	14	46.7
1,040	Max-Planck-Institut fur Mathematik in den Naturwissenschaften (MIS)	126	182	56	44.4
1,229	Thermo Fisher Scientific Inc.	23	33	10	43.5
140	FernUniversitat in Hagen	98	140	42	42.9
635	Hochschule Esslingen	26	37	11	42.3
645	Hochschule Darmstadt	43	61	18	41.9
145	Padagogische Hochschule Freiburg	24	34	10	41.7
4,123	Bremer Institut fur Praventionsforschung und Sozialmedizin (BIPS)	101	143	42	41.6
5,210	Berliner Institut fur Gesundheitsforschung	805	1,139	334	41.5
1,614	AUDI AG	22	31	9	40.9

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
603	Hochschule fur Technik, Wirtschaft und Kultur Leipzig	27	38	11	40.7

Table 16: Institutions with more than 20 publications in 2018 in the wos_b_2019 that decreased in publication counts by over 40% to 2019 in the wos_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
1,637	Zentrum fur Rhinologie und Allergologie	44	26	-18	-40.9
756	Forschungszentrum caesar	44	25	-19	-43.2
242	Europaische Kommission Gemeinsame Forschungsstelle Institut fur Transurane	65	36	-29	-44.6
1,129	Fraunhofer-Institut fur Werkstoffmechanik	46	25	-21	-45.7
322	Klinikum Nurnberg	30	15	-15	-50.0
4,704	Schon Klinik Verwaltung GmbH	46	23	-23	-50.0
1,214	UCB Pharma GmbH	25	12	-13	-52.0
664	Hochschule fur angewandte Wissenschaften - Fachhochschule Aschaffenburg	30	13	-17	-56.7
1,226	ThyssenKrupp AG	37	15	-22	-59.5
1,458	Henkel AG	22	8	-14	-63.6
24	Kiepenheuer-Institut fur Sonnenphysik (KIS)	35	12	-23	-65.7
593	Rheinische Fachhochschule Koln	29	6	-23	-79.3
4,618	Paul Gerhardt Diakonie	22	4	-18	-81.8

Authors: Mean number of authors by Research Area and discipline

The mean number of authors on a paper can be informative about patterns of collaboration and their potential implications for fractional counting. For instance, increasing levels of inter-sector or international collaboration could result in decreased publication counts for individual sectors or countries when using fractional counting. As such, understanding changes in authorship patterns can provide some insight into potential macro-level changes for entities.

We show in the left panel of Figure 15 the mean number of authors per sc_extended discipline in 2018 in both databases, and in the right panel the mean number of authors per discipline in 2018 in the wos_b_2020 database compared to 2019 in the wos_b_2020 database.

While little change is expected to be seen in the left-hand panel of Figure 15 as the number of authors on a paper is unlikely to change between databases, differences in the right-hand panel indicate potential changes in disciplines' collaboration patterns. Disciplines for which the mean number of authors changed by more than 5%, based on the right-hand panel of Figure 15, are shown in Table 17. Also, to assess trends over a longer time-series, we present the mean number of authors per Research Area over the last ten common years of both databases in Figure 16.

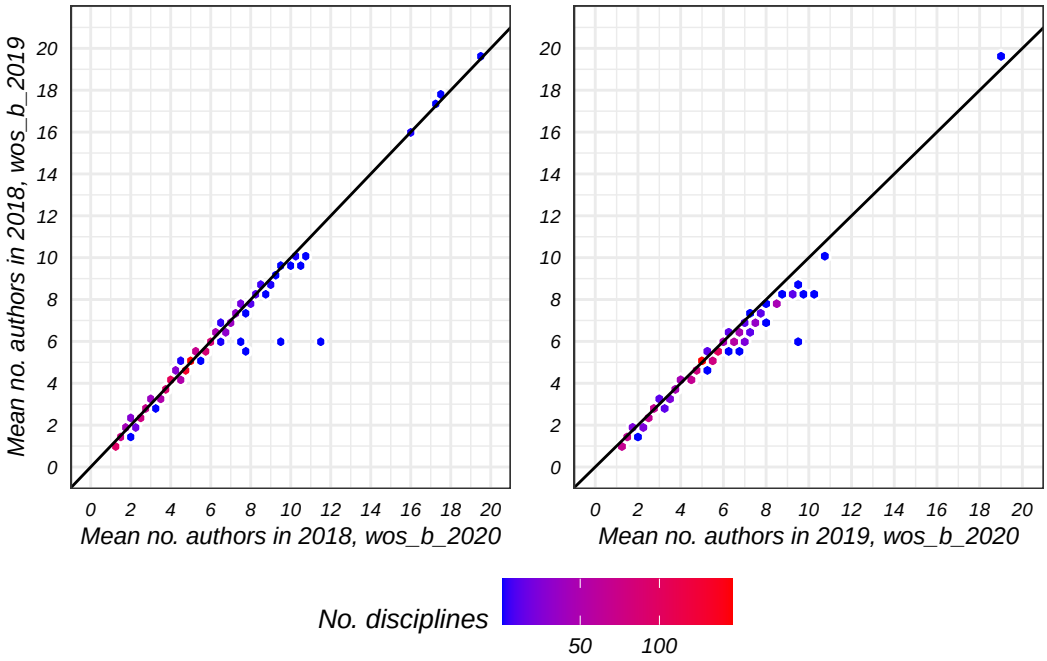

Figure 15: Mean number of authors per discipline (sc_extended) between databases, where colour denotes the number of disciplines with this combination of mean authors.

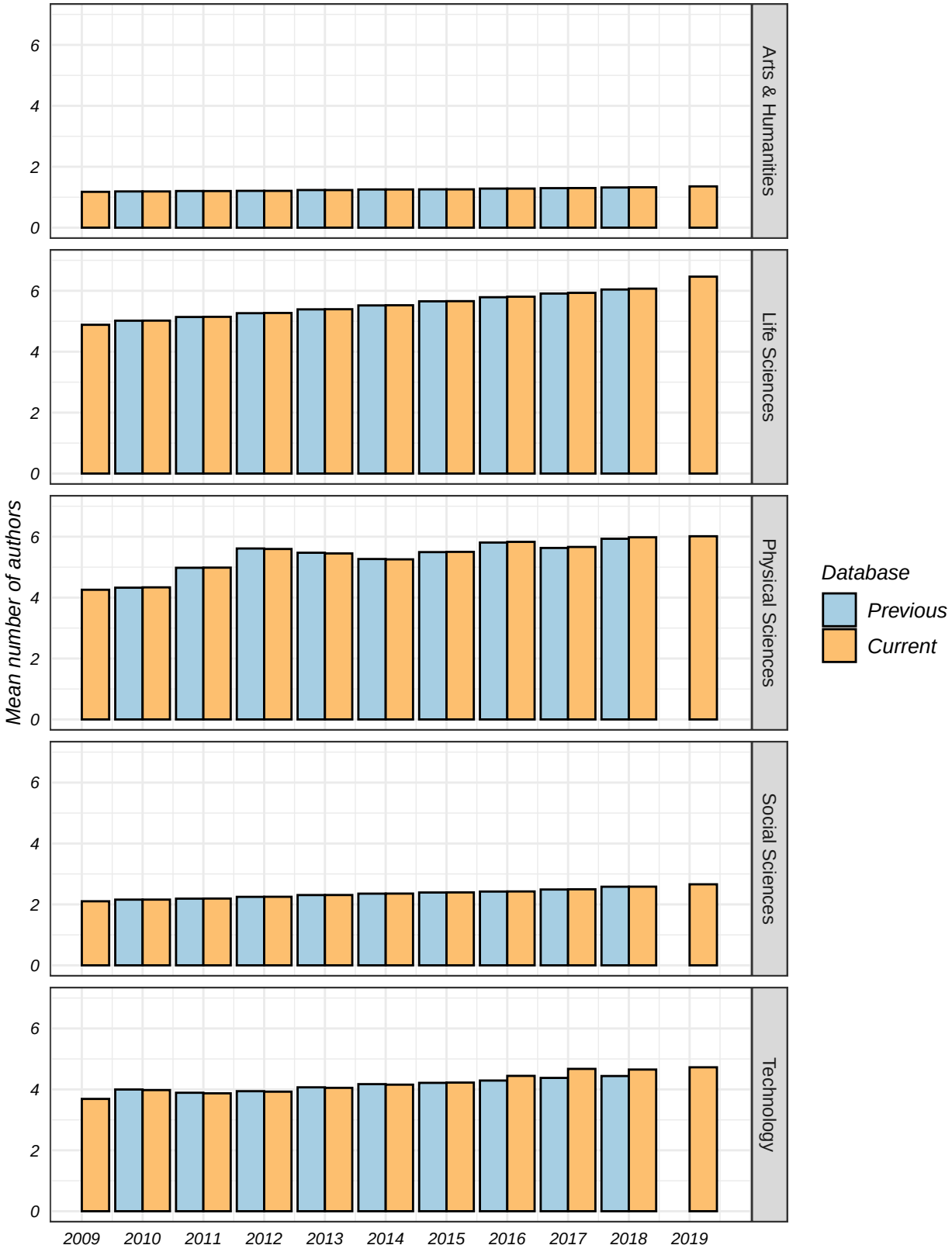

Figure 16: Mean number of authors by Research Area and database over time.

Table 17: Disciplines (sc_extended) where the mean number of authors changed by more than 10% between the last common year in the previous database and the latest year in the current database.

Discipline	Previous mean authors	Current mean authors	Perc. diff.	No. crrnt pubs
Dance	1.2	1.1	-9.1	271
Area Studies	1.5	1.4	-7.1	2,950
Microscopy	5.1	4.9	-4.1	1,211
Women's Studies	2.8	2.7	-3.7	2,346
Astronomy & Astrophysics	19.5	19.0	-2.6	22,079
Geriatrics & Gerontology	6.1	6.8	10.3	9,152
General & Internal Medicine	6.9	7.7	10.4	40,073
Allergy	7.7	8.6	10.5	2,886
Gastroenterology & Hepatology	8.4	9.4	10.6	11,591
Emergency Medicine	5.8	6.5	10.8	4,025
Neurosciences & Neurology	6.5	7.3	11.0	65,092
Architecture	1.6	1.8	11.1	1,952
Archaeology	3.1	3.5	11.4	3,490
Respiratory System	7.6	8.7	12.6	9,570
Rheumatology	7.6	8.7	12.6	4,846
Pediatrics	5.9	6.8	13.2	17,846
Anesthesiology	6.0	7.0	14.3	4,403
Infectious Diseases	8.1	9.6	15.6	16,439
Ophthalmology	5.7	6.9	17.4	9,514
Cardiovascular System & Cardiology	8.3	10.2	18.6	28,351
Nuclear Science & Technology	6.0	9.3	35.5	9,150

Source items: Percentage by Research Area and discipline

Source items refer to whether the publications on the reference list of an indexed publication are also indexed in the database, as opposed to not indexed and therefore non-source. Only source items are included in citation counts and so understanding the percentage of items cited that are also source can give an indication of the depth of WoS' coverage of a discipline. That is, if a large number of indexed items' sources are not indexed, the reverse is also likely true and a large number of citations of indexed items are also missing, which has the effect of reducing citation counts for disciplines with lower coverage, such as the arts and humanities.

The percentage of references that are source items is expected to increase over time as Clarivate continues to index journals and makes efforts to improve coverage of journals from disciplines with known low coverage. The percentage is not likely to ever reach 100% however, as authors will continue to cite items outside of the scope or coverage of WoS.

We show in the left-hand panel of Figure 17 the percentage of references that are source items per sc_extended discipline in 2018 in both databases, and in the right-hand panel the percentage of references that are source items per discipline in 2018 in the wos_b_2020 database compared to 2019 in the wos_b_2020 database.

It is in the right-hand panel that the effect of recently indexed journals may become apparent, where an increase in the percentage of source items may be seen if the journal is often cited within a discipline. The disciplines with a change in the percentage of indexed references of more than five percentage points between databases, based on the right-hand panel of Figure 17, are shown in Table 18. Longer term trends can be seen in Figure 18 where we present the percentage of reference that are source items per Research Area over the last ten common years of both databases.

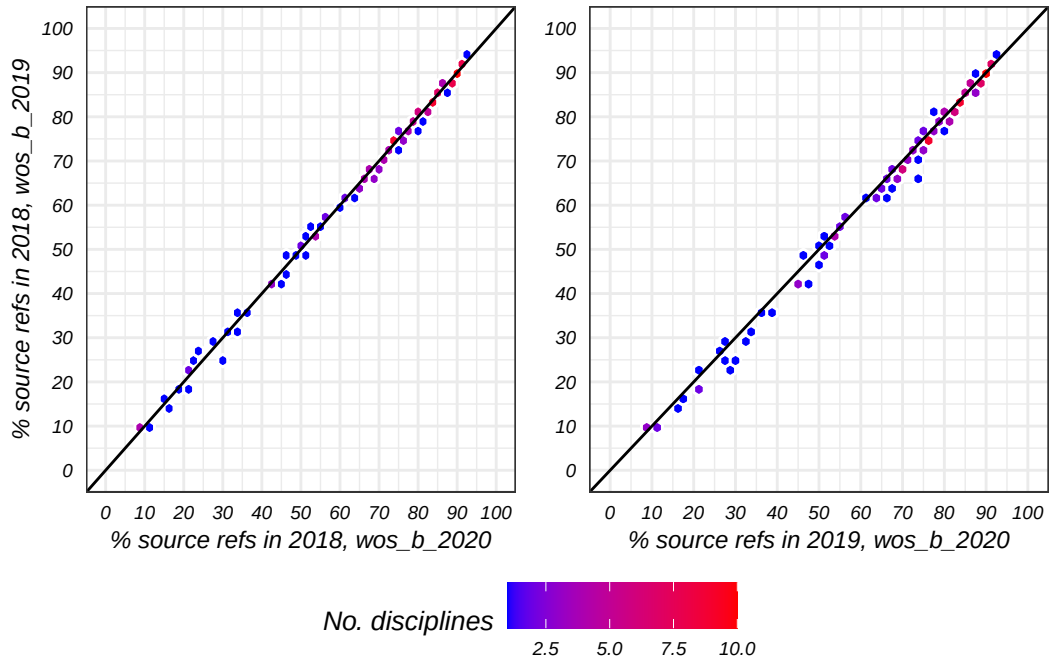

Figure 17: The percentage of cited items that are source items per sc_extended discipline by database, where colour denotes the number of disciplines with this combination of source references.

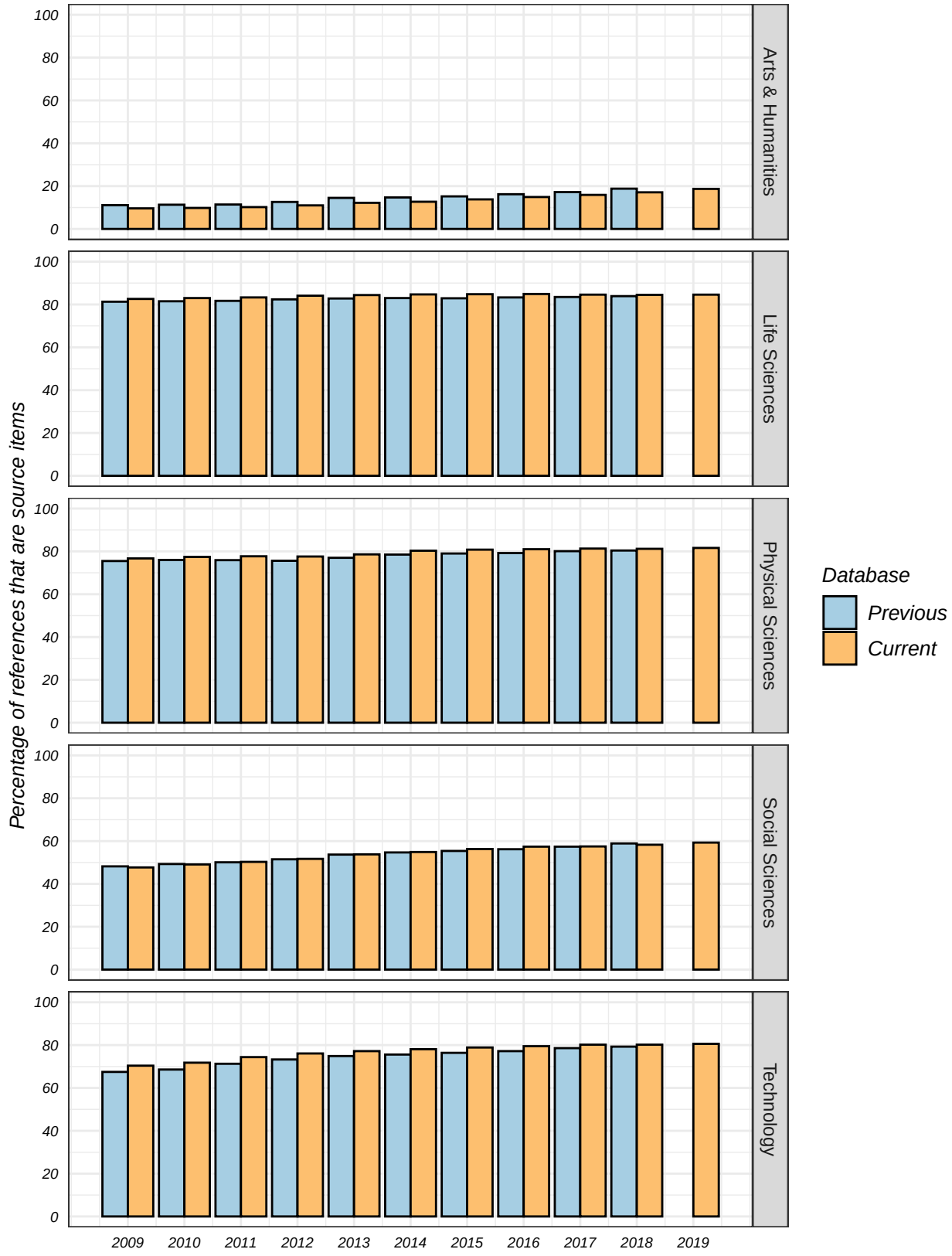

Figure 18: The percentage of references that are source items by Research Area and database over time.

Table 18: Disciplines (sc_extended) where the percentage of indexed references changed by more than 5 percentage points between 2018 in wos_b_2019 and 2019 in wos_b_2020.

Discipline	Previous no. refs.	Prvs % source	Current no. refs.	Crrnt % source	Change
Nuclear Science & Technology	2,104,049	66.9	2,937,018	74.1	7.2
Art	305,140	25.4	264,916	31.1	5.7
Communication	867,743	42.9	949,333	48.3	5.4
Film, Radio & Television	94,284	22.9	84,934	28.1	5.2
Social Issues	440,852	45.4	441,400	50.5	5.1

References

[1] S. Stahlschmidt, D. Stephen and S. Hinze. “Performance and Structures of the German Science System”. In: Studien zum deutschen Innovationssystem. Expertenkommission Forschung und Innovation (EFI), 2019. Chap. Studie 5-2019.

[2] J. Wang. “Citation time window choice for research impact evaluation”. In: *Scientometrics* 94.3 (2013). doi:10.1007/s11192-012-0775-9, pp. 851–872.