Dimity Stephen / Stephan Stahlschmidt / Paul Donner

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

Report on scopus_b_2019 and scopus_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 snapshot
- 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 scopus_b_2019 and scopus_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 scopus_b_2019 and scopus_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 scopus_b_2019 database against 2017 in the scopus_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-L?nder (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 All Science Journal Classification (ASJC) or the Subject Area classification. The ASJC is the fine-grain classification more commonly used in analyses by the DZHW. However, given it contains over 250 categories, it is sometimes useful to use a coarse-grain approach to present an overview of the disciplines. As such, for some indicators we present data on the Subject Area classification, which collapses the disciplines into 4 broad groups: Health Sciences, Life Sciences, Physical Sciences, and Social Sciences and Humanities.

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 scopus_b_2020 database contains a higher number of publications for the most recent years than the scopus_b_2019 database due to the continued indexing of items by Elsevier 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 Scopus' 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 Scopus-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.

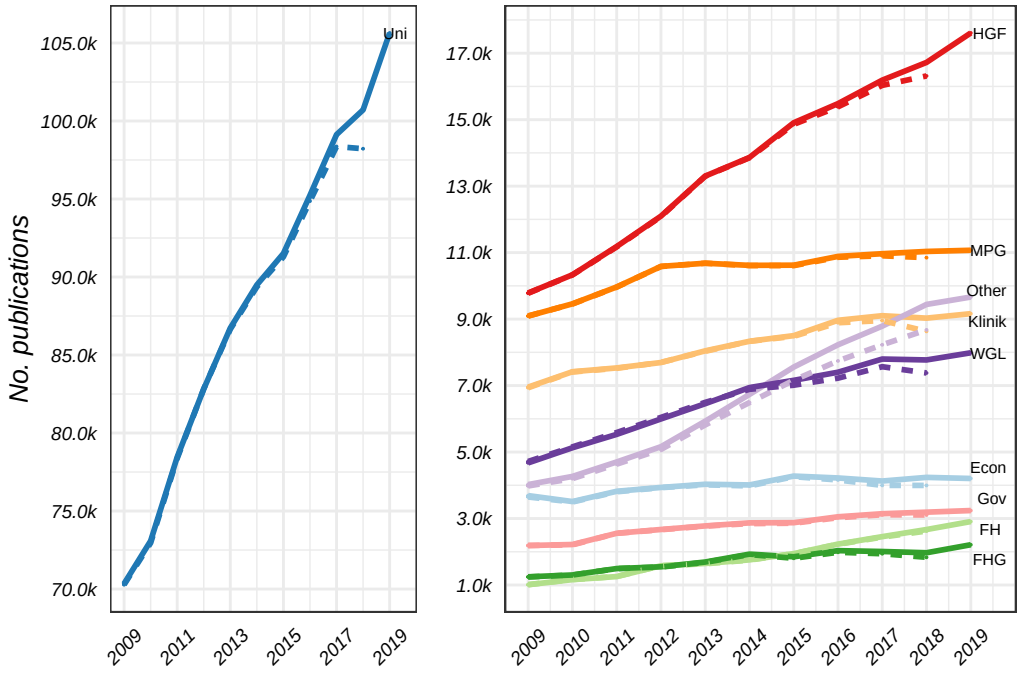

Figure 1: Whole counts of sectoral publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panel's different scales.

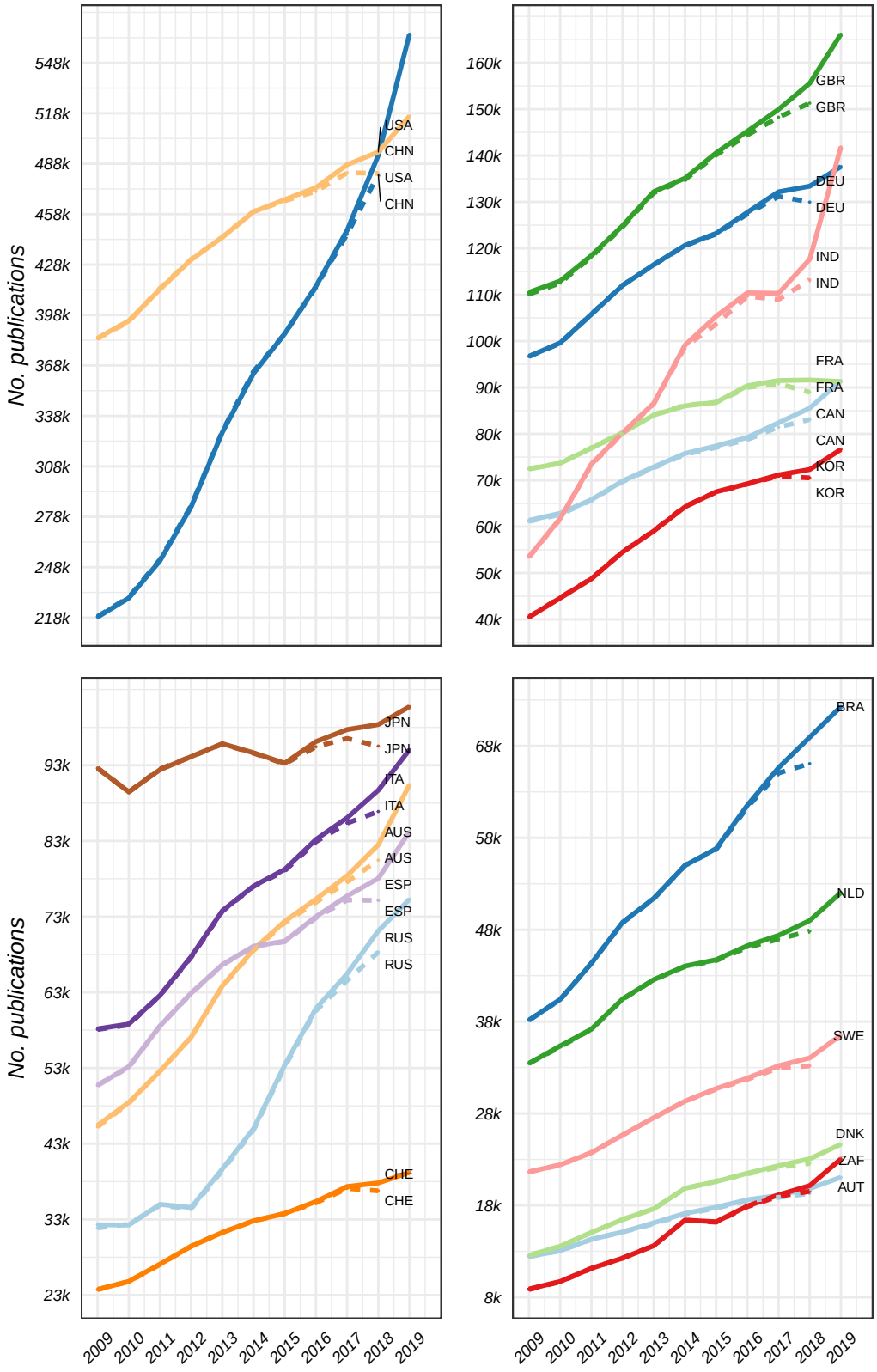

Figure 2: Whole counts of national publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels' different scales.

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 ASJC 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.

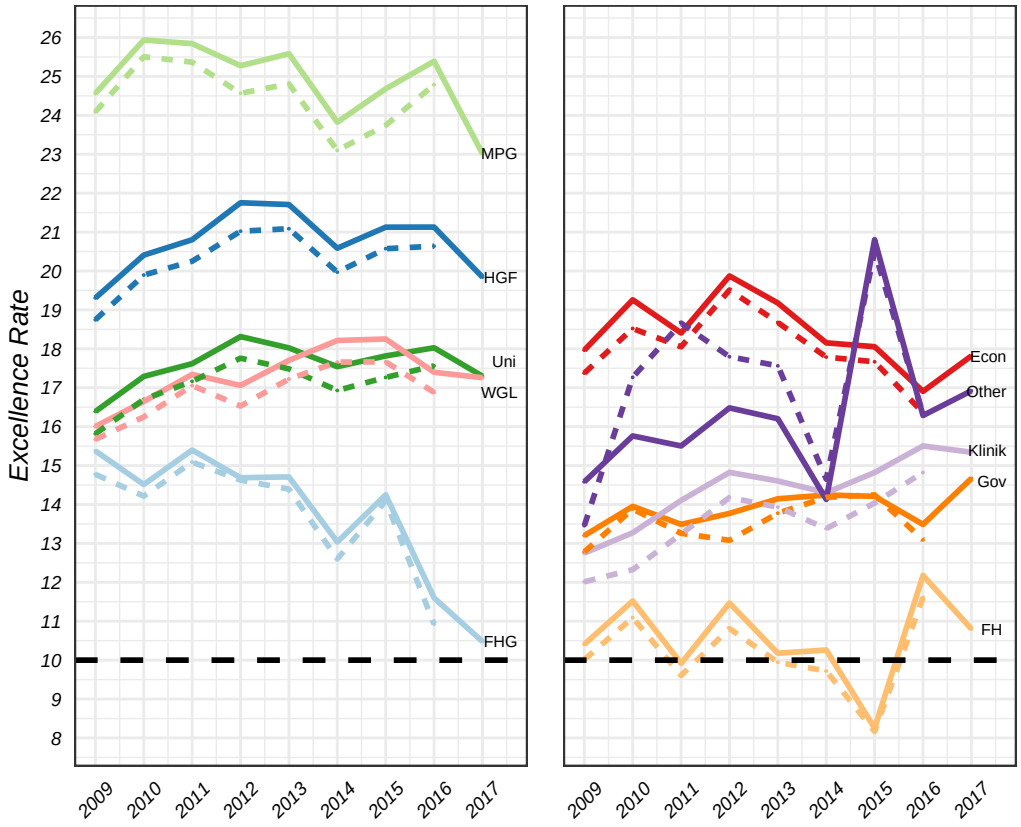

Figure 3: Excellence rates by sector, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

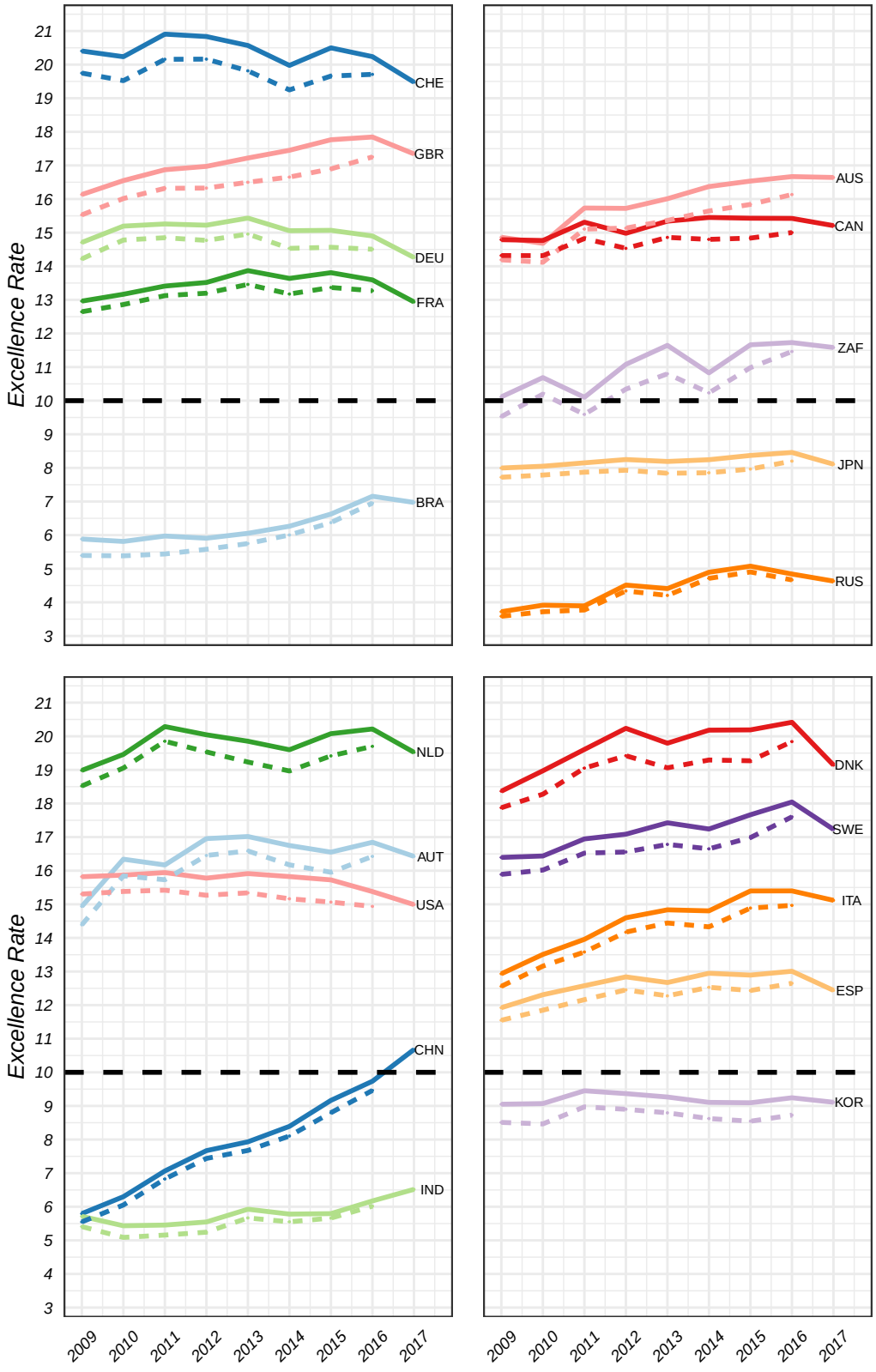

Figure 4: Excellence rates for selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

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 Scopus, compared to between Scopus and WoS, means we expect less change in the ER thresholds between the iterations of the Scopus 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 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 scopus_b_2019 and scopus_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 scopus_b_2019 database but now compared against the threshold in 2017 in the scopus_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.

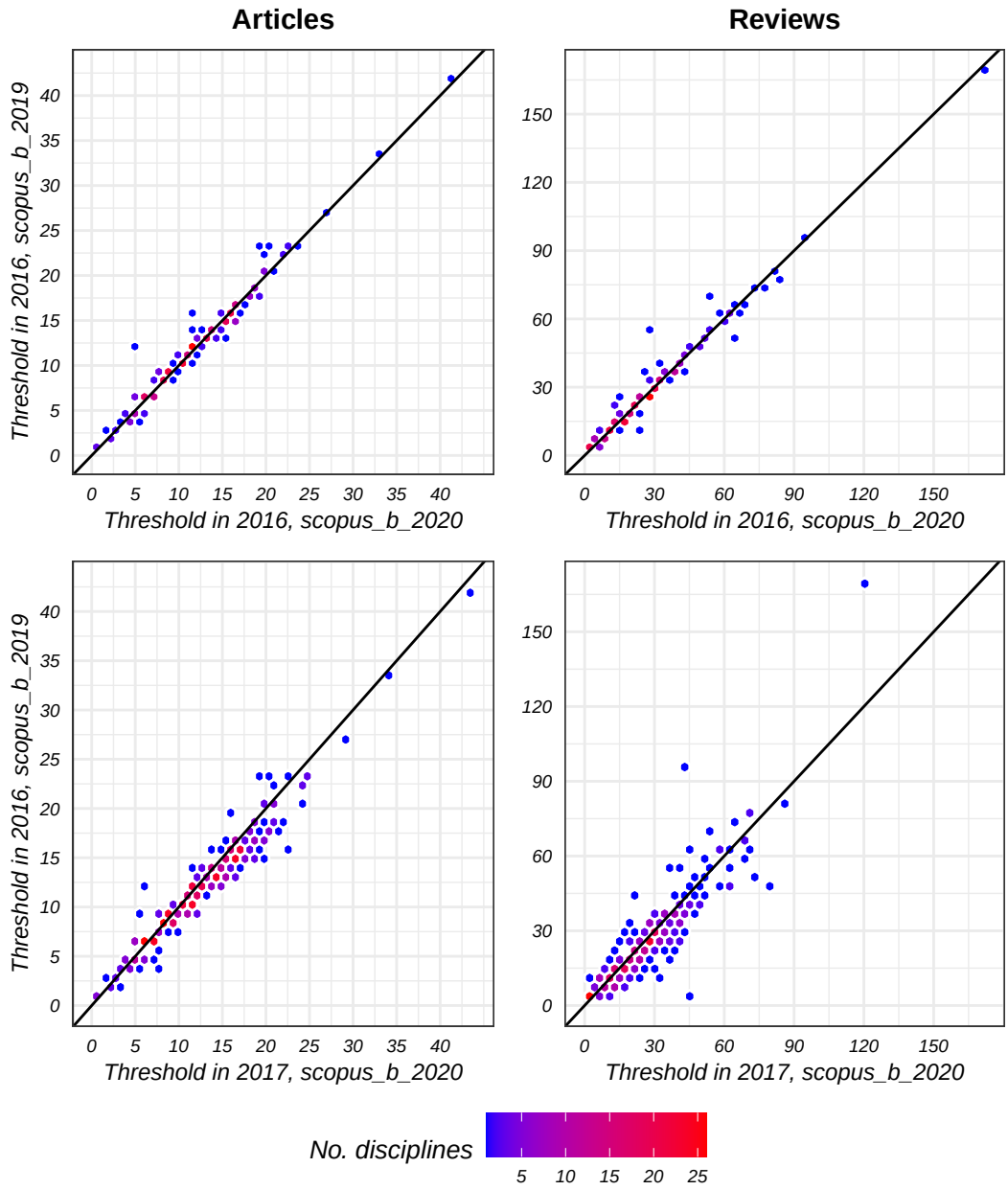

Figure 5: The ER threshold for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of thresholds.

Table 1: Articles: Disciplines where the ER threshold decreased by over 20% or increased by over 40% between 2016 in scopus_b_2019 and 2017 in scopus_b_2020 or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Pharmacology (nursing)	4	8	239	100.0
Architecture	4	6	4,196	50.0
Medical Assisting and Transcription	2	3	49	50.0
Neuroscience (miscellaneous)	14	11	1,361	-21.4
History	4	3	15,900	-25.0
Music	4	3	1,851	-25.0
Care Planning	7	5	324	-28.6
Decision Sciences (miscellaneous)	9	6	98	-33.3
Nurse Assisting	3	2	155	-33.3
General Medicine	12	6	52,529	-50.0

Table 2: Reviews: Disciplines with a current ER threshold of at least 10 where the ER threshold decreased by over 25% or increased by over 60% between 2016 in scopus_b_2019 and 2017 in scopus_b_2020 or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Discrete Mathematics and Combinatorics	4	46	10	1,050.0
Logic	10	33	10	230.0
Reviews and References (medical)	6	17	2	183.3
LPN and LVN	4	11	48	175.0
Chiropractics	4	10	41	150.0
Pharmacology, Toxicology and Pharmaceutics (miscellaneous)	11	23	302	109.1
Computational Mathematics	19	37	143	94.7
Mathematical Physics	16	30	189	87.5
Algebra and Number Theory	6	11	23	83.3
Family Practice	6	11	387	83.3
Organizational Behavior and Human Resource Management	9	16	450	77.8
Physics and Astronomy (miscellaneous)	21	37	112	76.2
Marketing	12	21	188	75.0
Human Factors and Ergonomics	15	25	70	66.7
Biochemistry, Genetics and Molecular Biology (miscellaneous)	49	81	210	65.3

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Computer Graphics and Computer-Aided Design	21	34	125	61.9
General Medicine	21	15	8,054	-28.6
Energy (miscellaneous)	169	119	188	-29.6
Statistical and Nonlinear Physics	54	38	126	-29.6
Filtration and Separation	62	43	135	-30.6
Economic Geology	30	20	91	-33.3
Gerontology	24	14	251	-41.7
Environmental Science (miscellaneous)	34	19	281	-44.1
Paleontology	20	11	133	-45.0
Computer Science (miscellaneous)	31	16	63	-48.4
Statistics, Probability and Uncertainty	44	21	111	-52.3
Nuclear Energy and Engineering	95	45	243	-52.6

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. 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.

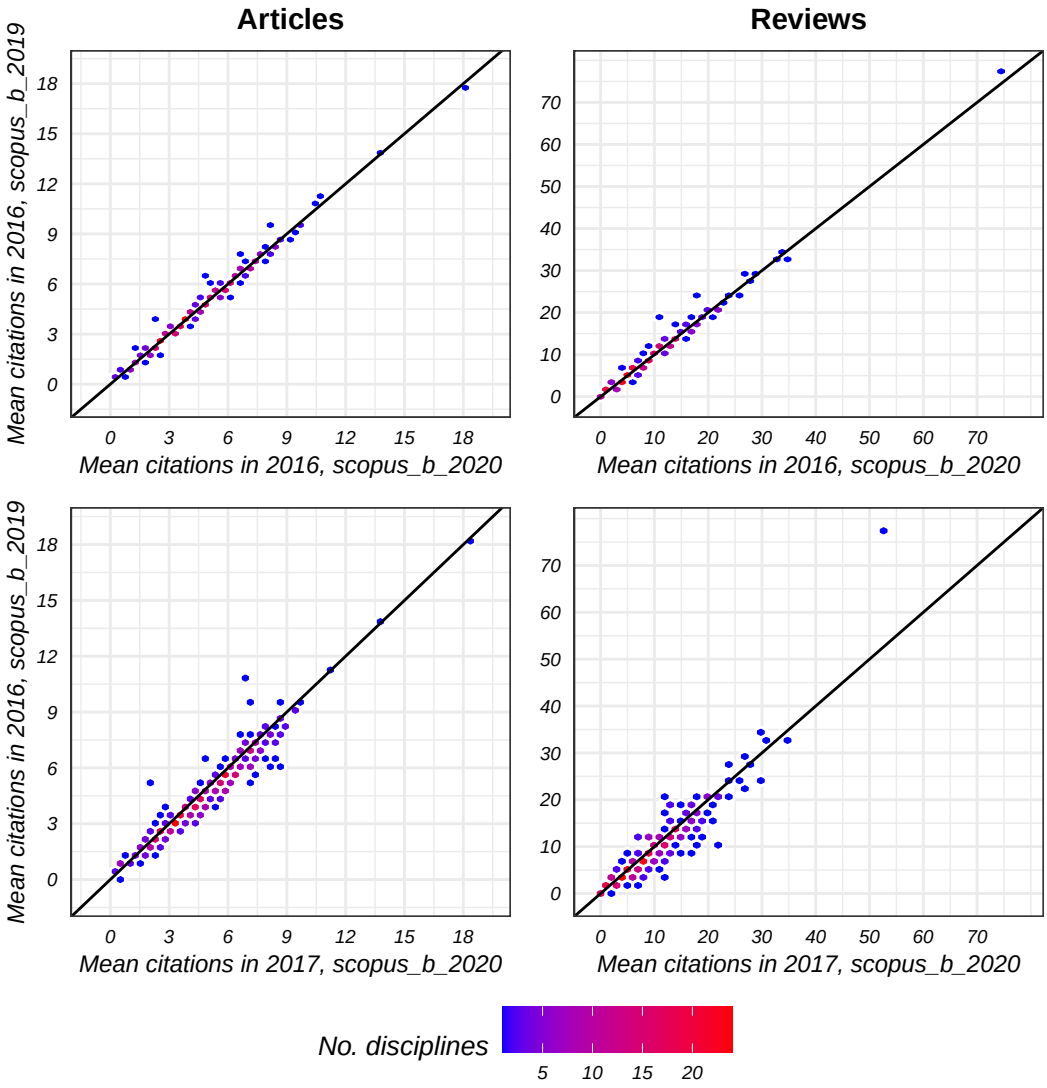

Figure 6: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

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 scopus_b_2019 and 2017 in scopus_b_2020, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
LPN and LVN	1.4	2.3	997	64.3
Pharmacology (nursing)	1.3	2.0	239	53.8
Care Planning	2.8	2.2	324	-21.4
Neuroscience (miscellaneous)	6.3	4.9	1,361	-22.2
Dental Hygiene	2.5	1.9	58	-24.0
General Medicine	4.0	3.0	52,529	-25.0
Hematology	9.4	7.0	11,229	-25.5
Decision Sciences (miscellaneous)	3.3	2.4	98	-27.3
Nurse Assisting	1.0	0.7	155	-30.0
Health Information Management	10.6	6.8	1,492	-35.8
Chiropractics	5.2	1.9	290	-63.5

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 scopus_b_2019 and 2017 in scopus_b_2020, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Discrete Mathematics and Combinatorics	1.6	5.1	10	218.7
Logic	2.1	6.2	10	195.2
LPN and LVN	1.3	3.7	48	184.6
Reviews and References (medical)	4.1	11.2	2	173.2
Computer Vision and Pattern Recognition	10.4	21.8	86	109.6
Psychology (miscellaneous)	1.9	3.9	46	105.3
Computers in Earth Sciences	5.9	11.5	82	94.9
Acoustics and Ultrasonics	9.8	18.5	185	88.8
Accounting	3.8	7.1	205	86.8
Chemical Engineering (miscellaneous)	8.4	15.4	155	83.3
Organizational Behavior and Human Resource Management	2.8	5.1	450	82.1
Computer Graphics and Computer-Aided Design	9.1	16.3	125	79.1
Pharmacology, Toxicology and Pharmaceutics (miscellaneous)	4.5	8.0	302	77.8

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Business and International Management	3.8	6.7	560	76.3
Veterinary (miscellaneous)	6.4	11.1	75	73.4
Marketing	4.2	7.2	188	71.4
Occupational Therapy	3.3	5.6	76	69.7
Radiological and Ultrasound Technology	6.5	10.9	345	67.7
Applied Mathematics	5.5	9.0	574	63.6
Numerical Analysis	3.7	6.0	24	62.2
Computational Mathematics	8.1	13.1	143	61.7
Family Practice	2.3	3.7	387	60.9
Information Systems and Management	16.5	12.9	126	-21.8
Dentistry (miscellaneous)	6.0	4.5	57	-25.0
Economic Geology	11.3	8.4	91	-25.7
General Medicine	6.6	4.9	8,054	-25.8
Social Sciences (miscellaneous)	4.7	3.4	729	-27.7
Astronomy and Astrophysics	17.7	12.4	525	-29.9
Energy (miscellaneous)	77.4	53.4	188	-31.0
Neuroscience (miscellaneous)	12.9	8.8	226	-31.8
Nuclear Energy and Engineering	18.2	12.3	243	-32.4
Instrumentation	19.7	13.3	665	-32.5
Gerontology	7.9	5.3	251	-32.9
General Decision Sciences	12.7	8.3	53	-34.6
Nuclear and High Energy Physics	20.1	12.6	269	-37.3
Environmental Science (miscellaneous)	11.7	6.7	281	-42.7
Computer Science (miscellaneous)	12.5	7.0	63	-44.0

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 Elsevier 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.

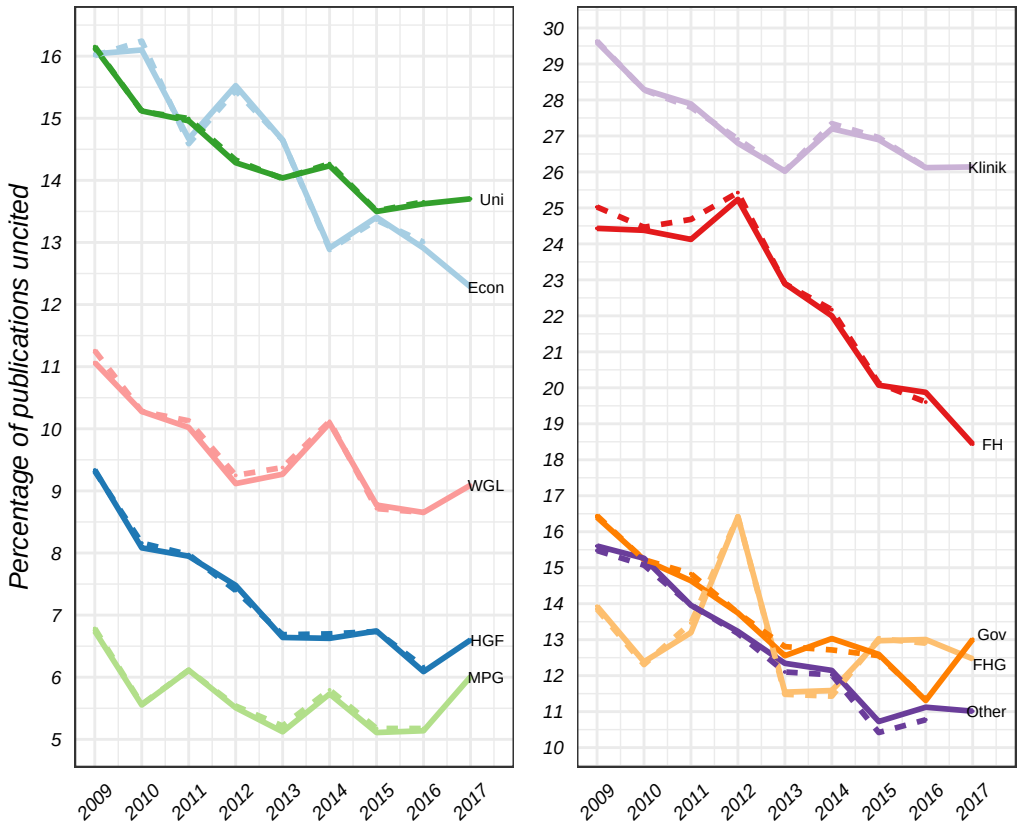

Figure 7: The percentage of uncited publications by German sector, based on whole counts, where dashed lines show the previous database and full lines show the current database.

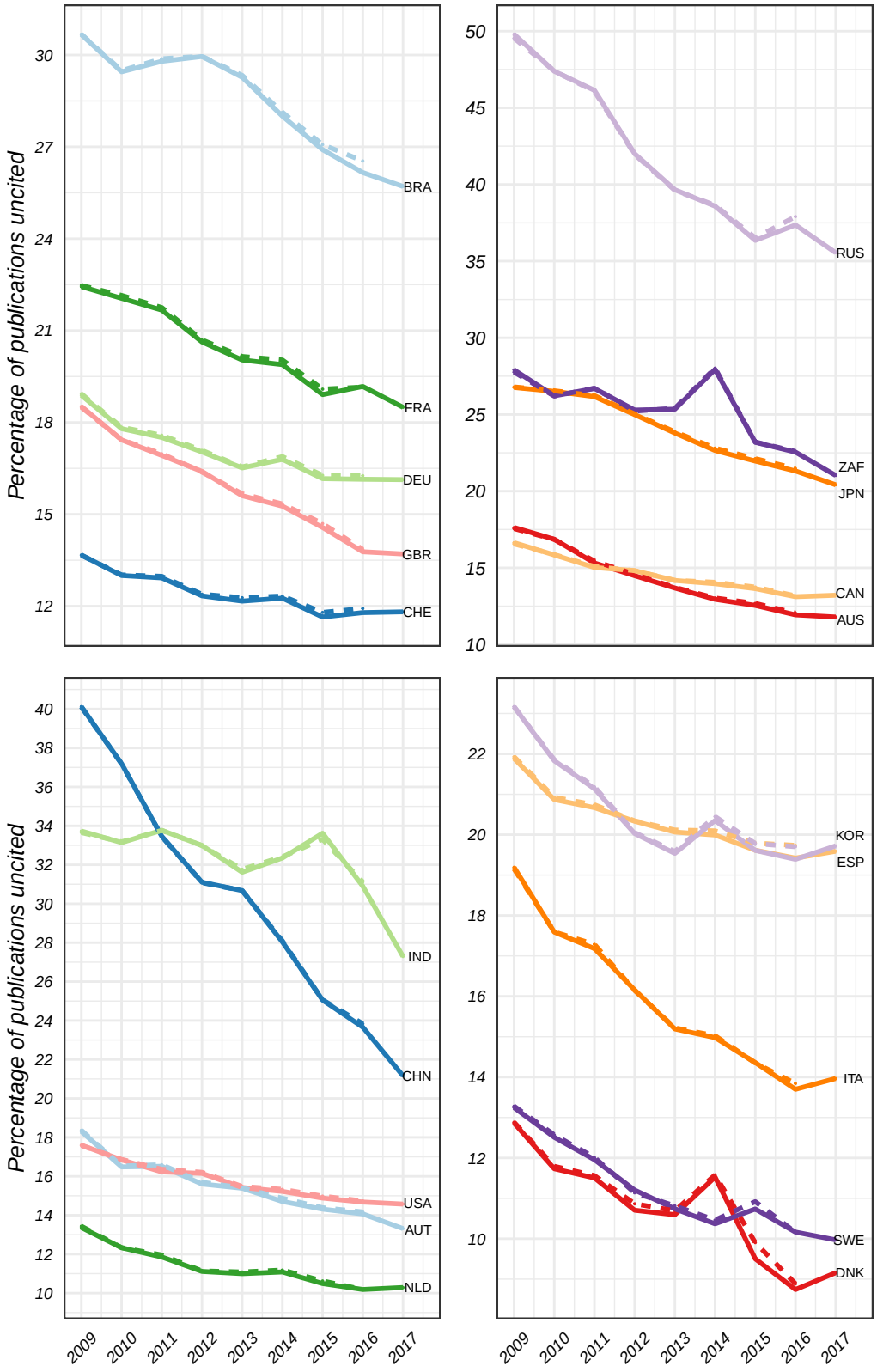

Figure 8: The percentage of uncited publications by selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database.

Disciplines: Changes in discipline classification

The two tables in this section highlight simply whether any changes have been made to Scopus' discipline classification, the All Science Journal Classification (AJSC). This could include splits, aggregations or removals of a discipline, or the inclusion of a new discipline to reflect new and emerging topics. Here we identify changes in the classification structure by comparing the number of articles and reviews attributed to each discipline in the latest year 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 Elsevier's published classification structure. Changes in the structure of the ASJC classification are shown in Table 5.

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

Code	Classification	Previous pubs	Current pubs
2744	Reviews and References (medical)	3	NA
3330	NA	NA	24

Disciplines: Changes in articles and reviews by discipline

This section examines 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 scopus_b_2019 and 2019 in scopus_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 ASJC. 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 scopus_b_2019 and 2019 in scopus_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.

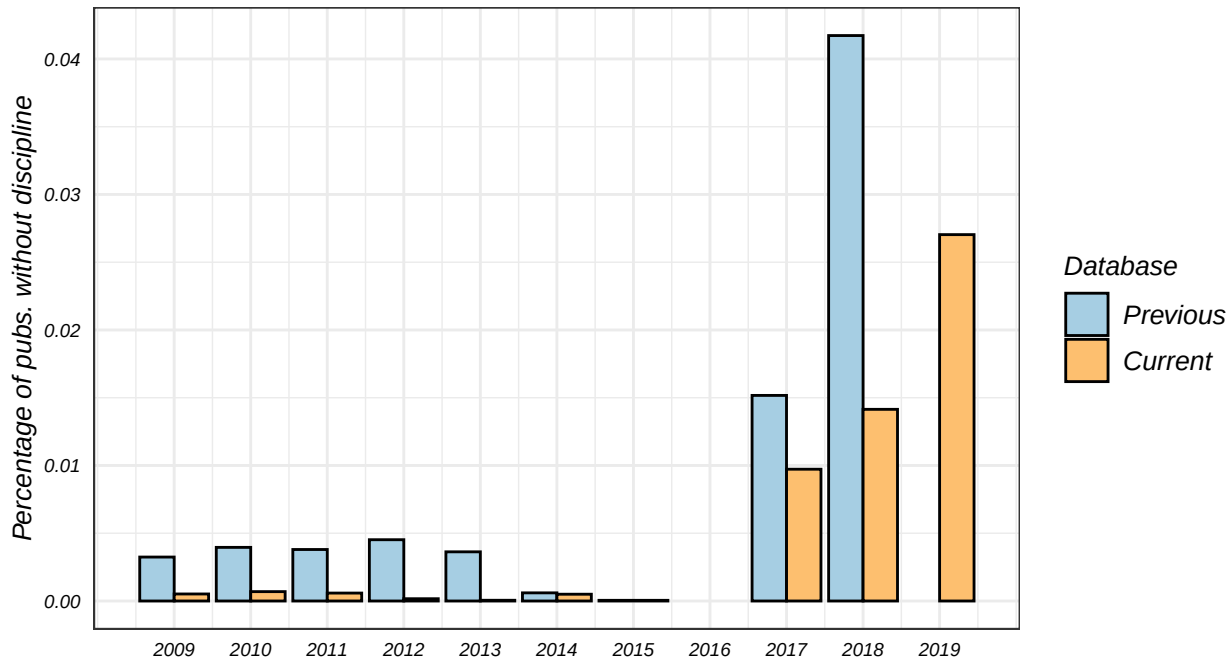

Figure 10: The percentage of publications not assigned to a discipline.

Metadata: Changes in pubyear, doctype and pubtype

This section details the number of items where 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 Elsevier 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 scopus_b_2019 and scopus_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 6 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 6: The number of items with changes in metadata between previous and current databases.

Year	Pub. year	Doc. type	Pub. type
2009	1461	438	159
2010	1118	533	112
2011	230	2735	1799
2012	300	1944	1029
2013	960	1146	76
2014	2114	1692	181
2015	1932	2255	94
2016	415	2151	59
2017	390	1628	97
2018	1604	6695	57

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 scopus_b_2019 are in Table 7 and scopus_b_2020 are in Table 8. 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 7: Annual percentage of publications missing metadata, scopus_b_2019.

Year	No page	No issue	No vol.	No DOI	No title	No refs	No abs.	No keys
2009	10.0	2.4	1.8	21.9	0	5.3	NA	NA
2010	10.2	2.7	1.5	21.0	0	5.1	NA	NA
2011	11.4	3.4	1.6	20.1	0	5.0	NA	NA
2012	13.0	7.8	1.5	18.0	0	4.4	NA	NA
2013	13.6	10.3	1.3	15.3	0	3.6	NA	NA
2014	13.2	13.0	1.2	13.8	0	4.1	NA	NA
2015	11.6	16.3	0.6	12.3	0	4.1	NA	NA
2016	14.0	18.5	0.6	10.4	0	3.4	NA	NA
2017	16.7	19.1	0.5	8.3	0	2.6	NA	NA
2018	18.3	20.0	0.6	6.6	0	3.1	NA	NA

Table 8: Annual percentage of publications missing metadata, scopus_b_2020.

Year	No page	No issue	No vol.	No DOI	No title	No refs	No abs.	No keys
2009	9.9	2.4	1.8	21.9	0	5.3	NA	NA
2010	10.2	2.7	1.5	21.0	0	5.1	NA	NA
2011	11.3	3.3	1.5	20.0	0	4.9	NA	NA
2012	13.0	7.7	1.5	17.9	0	4.3	NA	NA
2013	13.6	10.2	1.3	15.4	0	3.6	NA	NA
2014	13.2	12.9	1.2	13.9	0	4.1	NA	NA
2015	11.6	16.2	0.6	12.5	0	4.0	NA	NA
2016	14.1	18.4	0.6	10.6	0	3.4	NA	NA
2017	16.8	19.0	0.5	8.5	0	2.7	NA	NA
2018	18.4	19.7	0.6	7.1	0	2.8	NA	NA
2019	28.4	24.8	5.8	5.4	0	3.4	NA	NA

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.

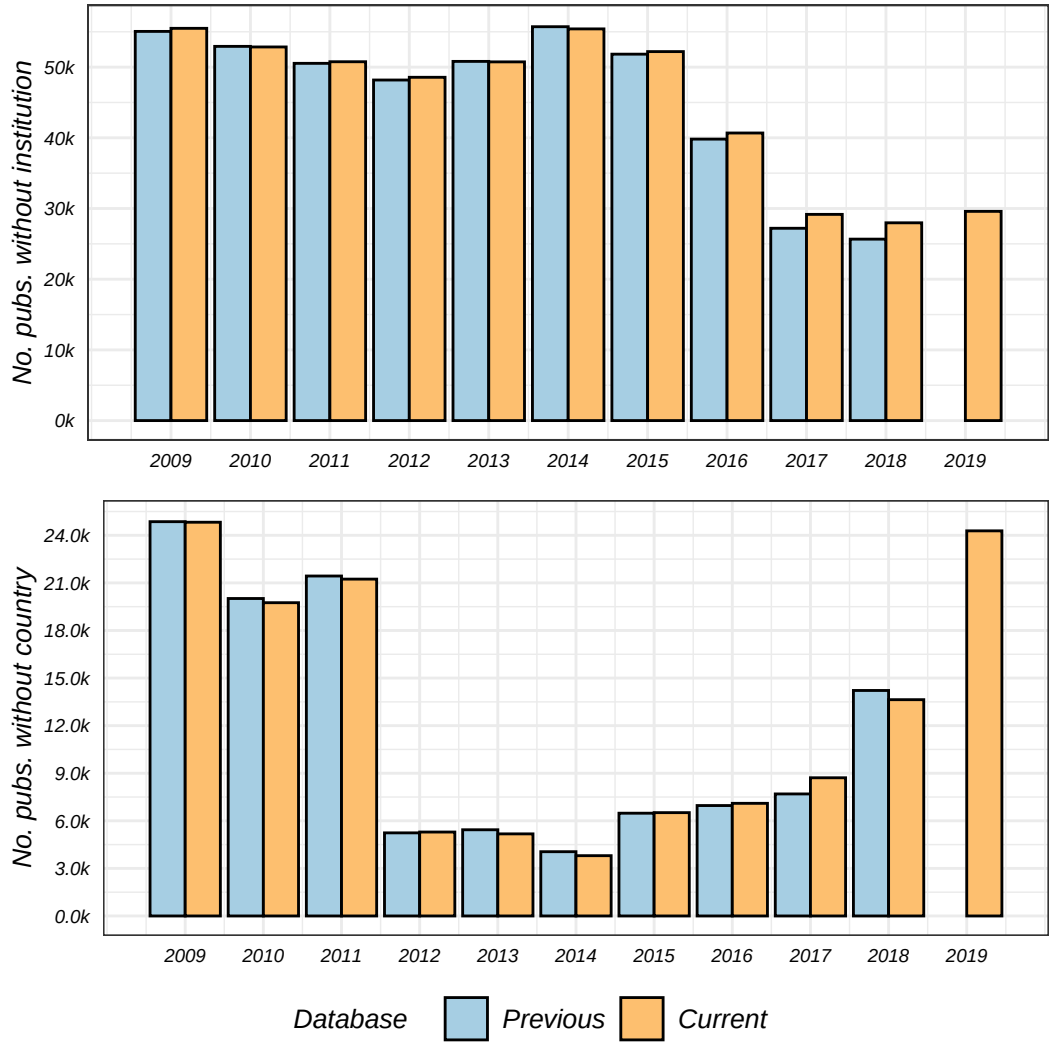

Figure 11: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

Author-institution links: Percentage complete by Subject 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 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 scopus_b_2019, now compared with the 2019 in the scopus_b_2020, indicating potential changes between the latest year in each database.

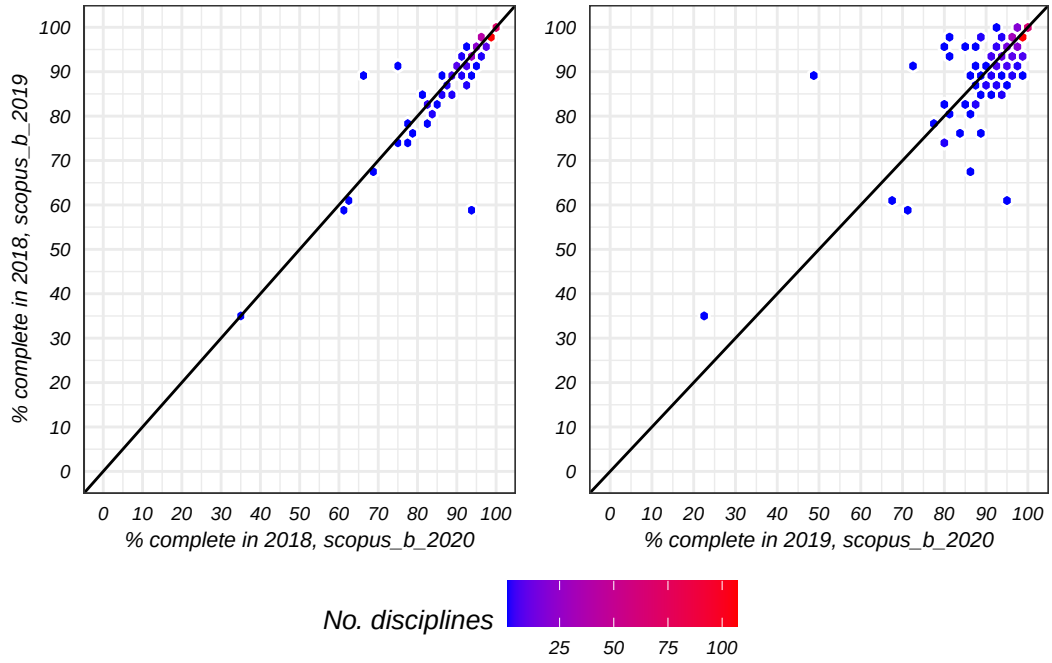

Figure 12: The percentage of complete author-institution links by disciplines.

The outlying disciplines observed in the right-hand panel of Figure 12 that have a change of more than 5 percentage points in the percentage of complete author-institution links between databases are shown in Table 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 Subject Area over the last ten common years in both databases. Substantial changes between years or differences between the databases may require investigation of the cause.

Table 9: Disciplines with a change of more than 7 percentage points in missing links between 2018 in scopus_b_2019 and 2019 in scopus_b_2020.

Discipline	Prvs items	% prvs complete	Crrnt items	% crrnt complete	Change
Medical Assisting and Transcription	350	312	534	255	-41.4
Medical Terminology	313	288	772	561	-19.3
Developmental Neuroscience	79,447	78,212	56,523	46,217	-16.7
Dental Hygiene	976	927	920	726	-16.1
Development	157,485	148,671	1,820,418	1,464,847	-13.9
Pharmacology (nursing)	5,189	1,820	7,875	1,824	-11.9
Psychology (miscellaneous)	16,731	15,878	25,119	21,484	-9.4
Experimental and Cognitive Psychology	166,345	162,111	204,806	180,743	-9.2
Hematology	327,407	318,202	864,264	766,474	-8.5
Social Psychology	139,506	133,523	269,244	235,803	-8.1
Fundamentals and Skills	11,838	10,306	14,103	13,317	7.4
Pediatrics	15,434	13,261	45,873	42,803	7.4
Embryology	15,623	13,850	18,521	17,805	7.5
Computer Vision and Pattern Recognition	250,053	224,406	289,245	284,130	8.5
Drug Guides	405	342	469	437	8.7
Veterinary (miscellaneous)	36,296	21,473	32,538	22,811	11.0
Decision Sciences (miscellaneous)	690	531	1,643	1,468	12.4
Medical and Surgical Nursing	11,246	7,574	35,076	30,284	19.0
Podiatry	3,452	2,064	4,903	4,666	35.4

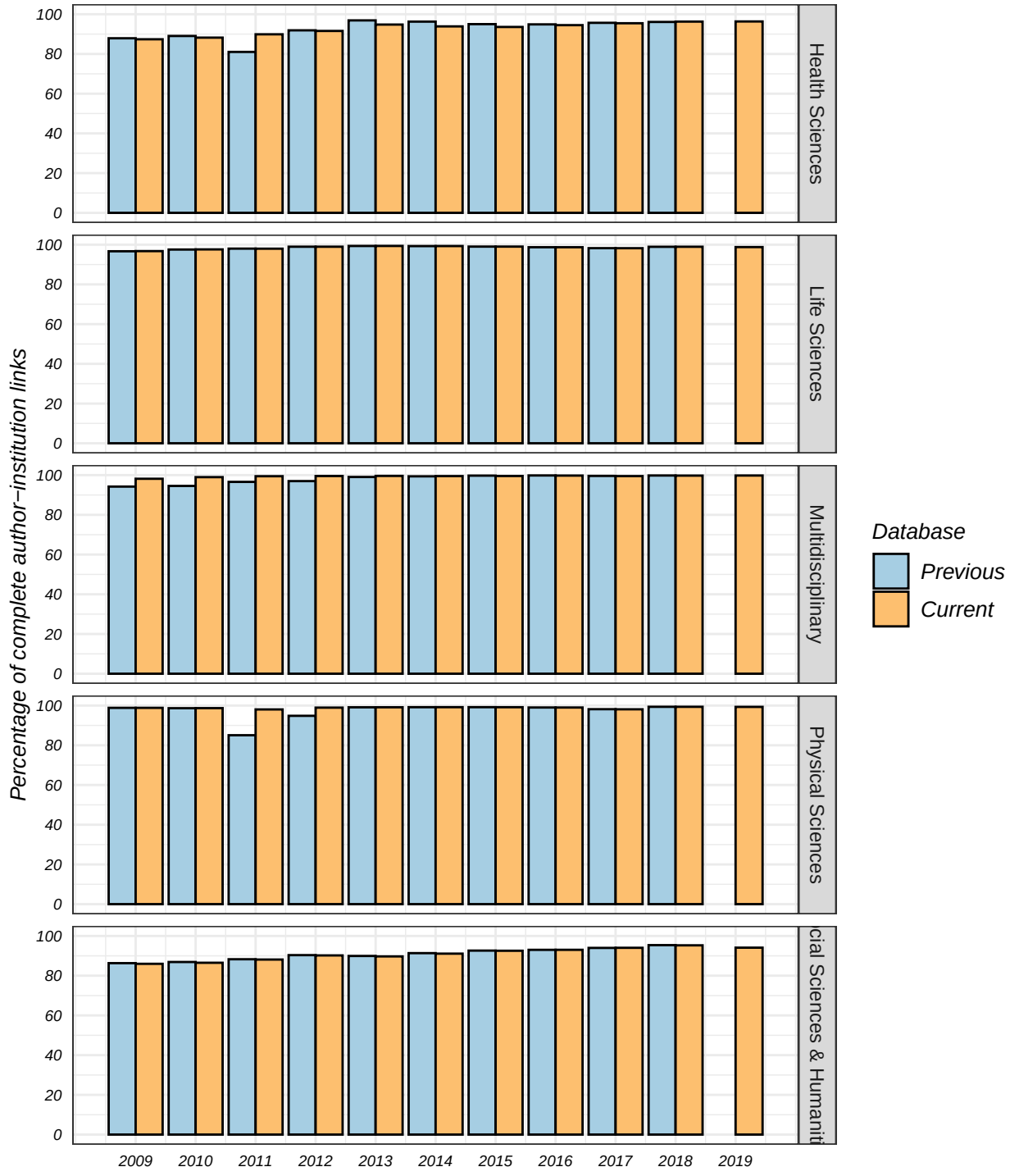

Figure 13: The annual percentage of complete author-institution links by Subject Area and database.

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.

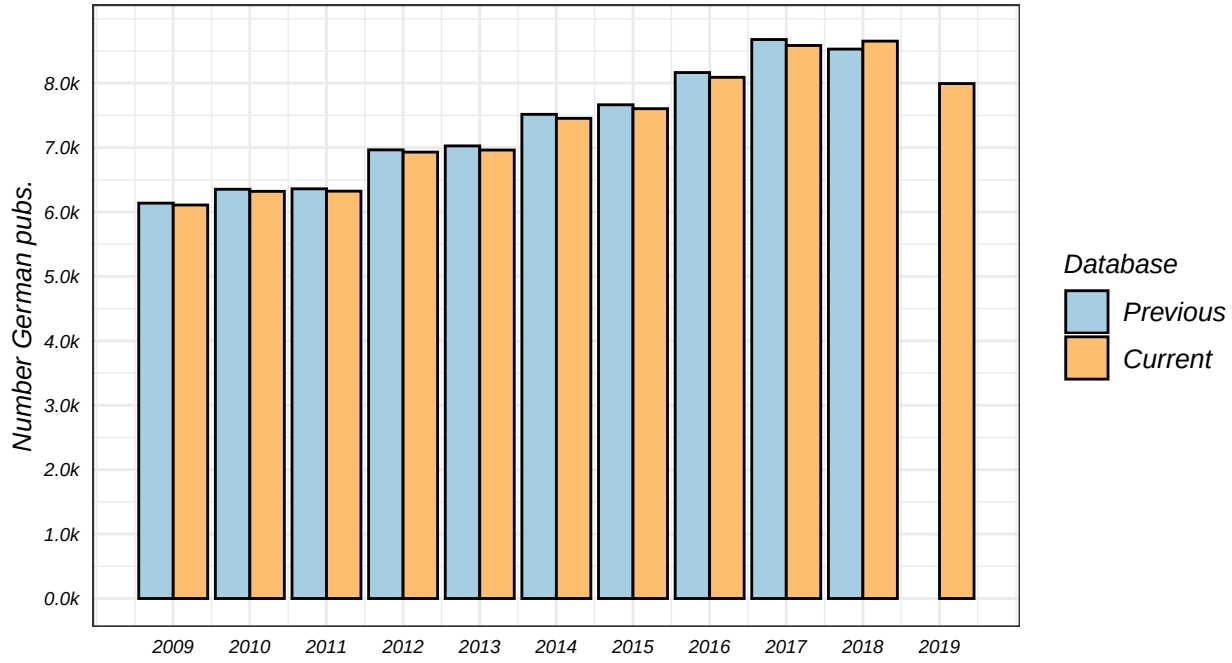

Figure 14: The number of German publications not assigned to a KB institution.

German institutions: Changes in whole counts of articles and reviews

This section compares changes in the number of articles and reviews published by German institutes between the latest years available in each database. These tables can assist in identifying institutions that have substantial numbers of publications which 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 10 presents potentially new institutions – these had no publications in 2018 in the scopus_b_2019 database but more than five publications in 2019 in the scopus_b_2020 database. Conversely, Table 11 shows the potentially decommissioned institutions that had at least five publications in 2018 in the scopus_b_2019 database but no publications recorded in 2019 in the scopus_b_2020 database. We also highlight in Tables 12 and 13 the larger institutions (with at least 20 publications) that increased or decreased their number of publications by more than 40% between 2018 and 2019 in the scopus_b_2019 and scopus_b_2020 databases.

Table 10: Institutions with more than 5 publications in 2019 in scopus_b_2020 that had no publications in 2018 in the scopus_b_2019 database.

PK_KB_INST	Name	Previous pubs	Current pubs
5477	Leibniz-Institut fur Photonische Technologien e.V. (IPHT)	0	181
5203	Nanosystems Initiative Munich (NIM)	0	74
5431	Nationales Centrum fur Tumorerkrankungen / University Cancer Center	0	49
175	International Psychoanalytic University Berlin	0	36
5488	Helmholtz-Institut fur Funktionelle Marine Biodiversitat an der Universitat Oldenburg	0	36
5478	Leibniz-Institut fur Werkstofforientierte Technologien - IWT	0	34
5432	Centogene AG	0	31
2121	Fraunhofer-Institut fur Mikrostruktur von Werkstoffen und Systemen	0	28
5483	Fraunhofer-Institut fur Energiewirtschaft und Energiesystemtechnik	0	23
1971	Technische Hochschule Koln	0	19
5479	Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)	0	19
5476	Leibniz-Institut fur Ost- und Sudosteuropaforschung	0	15
5482	Leibniz-Zentrum Moderner Orient (ZMO)	0	11
5484	Fraunhofer-Einrichtung fur Additive Produktionstechnologien IAPT	0	8
500	Bezirkskrankenhaus Augsburg Klinik fur Psychiatrie, Psychotherapie und Psychosomatik	0	6
5430	IKDT - INSTITUT KARDIALE DIAGNOSTIK und THERAPIE GMBH	0	6
5433	Deutsche Gesellschaft fur Gynakologie und Geburtshilfe e.V.	0	6
5438	BUCERIUS LAW SCHOOL	0	6
5510	Ingress-Health HWM GmbH	0	6

Table 11: Institutions with no publications in 2019 in scopus_b_2020 that had more than 5 publications in 2018 in the scopus_b_2019 database.

PK_KB_INST	Name	Previous pubs	Current pubs
3989	International Max-Planck Research School on Astrophysics	37	0
5277	Hochschule fur angewandte Wissenschaften Wurzburg Schweinfurt	12	0
3769	Berliner Wasserbetriebe	8	0
1351	microfluidic ChipShop GmbH	6	0
4172	Bernstein Fokus: Neurotechnologie (BFNT)	6	0
4552	Horzentrum Oldenburg GmbH	6	0

Table 12: Institutions with more than 20 publications in 2018 in the scopus_b_2019 that increased in publication counts by over 40% to 2019 in the scopus_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs	Perc. diff.
35	Leibniz-Institut fur Oberflachenmodifizierung (IOM)	25	87	248.0
1156	Fraunhofer-Institut fur Integrierte Schaltungen (IIS)	24	72	200.0
5290	Deutsches Zentrum fur Infektionsforschung	193	539	179.3
523	Klinikum Bayreuth GmbH	26	50	92.3
492	Klinikum Bremen-Mitte gGmbH	31	59	90.3
8	Wissenschaftszentrum Berlin fur Sozialforschung (WZB)	42	79	88.1
1146	Fraunhofer-Institut fur Produktionstechnik und Automatisierung	22	41	86.4
646	Hochschule fur angewandte Wissenschaften Coburg	21	39	85.7
4396	Hochschule Rhein-Waal - University of Applied Sciences	25	46	84.0
460	St.-Johannes-Hospital Dortmund	24	43	79.2
1140	Fraunhofer-Institut fur Silicatforschung (ISC)	50	87	74.0
561	Hochschule RheinMain, RheinMain University of Applied Sciences Wiesbaden, Russelsheim	21	36	71.4
604	Hochschule fur Wirtschaft und Recht Berlin	24	41	70.8
4696	European XFEL GmbH	65	107	64.6
1127	Fraunhofer-Institut fur Werkzeugmaschinen und Umformtechnik	42	69	64.3
1151	Fraunhofer-Institut fur Molekularbiologie und Angewandte Okologie	118	193	63.6

PK_KB_INST	Name	Previous pubs	Current pubs	Perc. diff.
643	Fachhochschule Dortmund	21	33	57.1
1040	Max-Planck-Institut fur Mathematik in den Naturwissenschaften (MIS)	128	200	56.2
145	Padagogische Hochschule Freiburg	25	39	56.0
56	Fraunhofer-Institut fur Optronik, Systemtechnik und Bildauswertung IOSB	27	42	55.6
552	Technische Hochschule Wildau (FH)	27	42	55.6
654	Fachhochschule Bielefeld	51	79	54.9
152	Universitat Erfurt	63	97	54.0
134	Hochschule fur Angewandte Wissenschaften Hamburg	77	117	51.9
586	Hochschule Magdeburg-Stendal	27	41	51.9
1021	Max-Planck-Institut fur Stoffwechselforschung	53	80	50.9
244	Europaische Zentralbank	79	118	49.4
648	Hochschule Bonn-Rhein-Sieg, University of Applied Sciences	43	64	48.8
635	Hochschule Esslingen	27	40	48.1
4123	Bremer Institut fur Praventionsforschung und Sozialmedizin (BIPS)	98	144	46.9
826	Deutsche Bundesbank	47	69	46.8
645	Hochschule Darmstadt	43	63	46.5
1103	GSI Helmholtzzentrum fur Schwerionenforschung	458	670	46.3
128	Padagogische Hochschule Heidelberg	25	36	44.0
33	Leibniz-Institut fur Pflanzenbiochemie	88	126	43.2
133	HafenCity Universitat Hamburg	21	30	42.9
450	Alfried Krupp von Bohlen und Halbach Krankenhaus gemeinnutzige GmbH	42	60	42.9
1069	Max-Planck-Institut fur Evolutionsbiologie	77	110	42.9
581	Hochschule fur angewandte Wissenschaften Munchen	64	91	42.2
153	Katholische Universitat Eichstatt - Ingolstadt	86	122	41.9
5283	Berlin-Brandenburgisches Institut fur Biodiversitatsforschung (BBIB)	73	103	41.1
804	Institut fur Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur fur Arbeit (BA)	61	86	41.0
477	Klinikum Darmstadt	37	52	40.5
330	St. Franziskus-Hospital GmbH Munster	52	73	40.4
5210	Berliner Institut fur Gesundheitsforschung	470	660	40.4

Table 13: Institutions with more than 20 publications in 2018 in the scopus_b_2019 that decreased in publication counts by over 40% to 2019 in the scopus_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs	Perc. diff.
1637	Zentrum fur Rhinologie und Allergologie	49	29	-40.8
322	Klinikum Nurnberg	45	26	-42.2
1214	UCB Pharma GmbH	26	15	-42.3
1226	ThyssenKrupp AG	35	20	-42.9
437	Krankenhaus Nordwest GmbH	30	17	-43.3
4743	Bioscientia - Institut fur Medizinische Diagnostik GmbH	26	14	-46.2
4704	Schon Klinik Verwaltung GmbH	44	23	-47.7
221	CeNTech GmbH - Center for Nanotechnology	27	14	-48.1
24	Kiepenheuer-Institut fur Sonnenphysik (KIS)	34	17	-50.0
756	Forschungszentrum caesar	44	22	-50.0
664	Hochschule fur angewandte Wissenschaften - Fachhochschule Aschaffenburg	34	14	-58.8
1491	Evonik Industries AG	22	9	-59.1
4618	Paul Gerhardt Diakonie	22	4	-81.8
593	Rheinische Fachhochschule Koln	60	3	-95.0

Authors: Mean number of authors by Subject 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 discipline in 2018 in both databases, and in the right panel the mean number of authors per discipline in 2018 in the scopus_b_2020 database compared to 2019 in the scopus_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. The disciplines with a more than 10% change in the mean number of authors based on the right-hand panel of Figure 15 are shown in Table 14. Also, to assess trends over a longer time-series, we present the mean number of authors per Subject Area over the last ten common years of both databases in Figure 16.

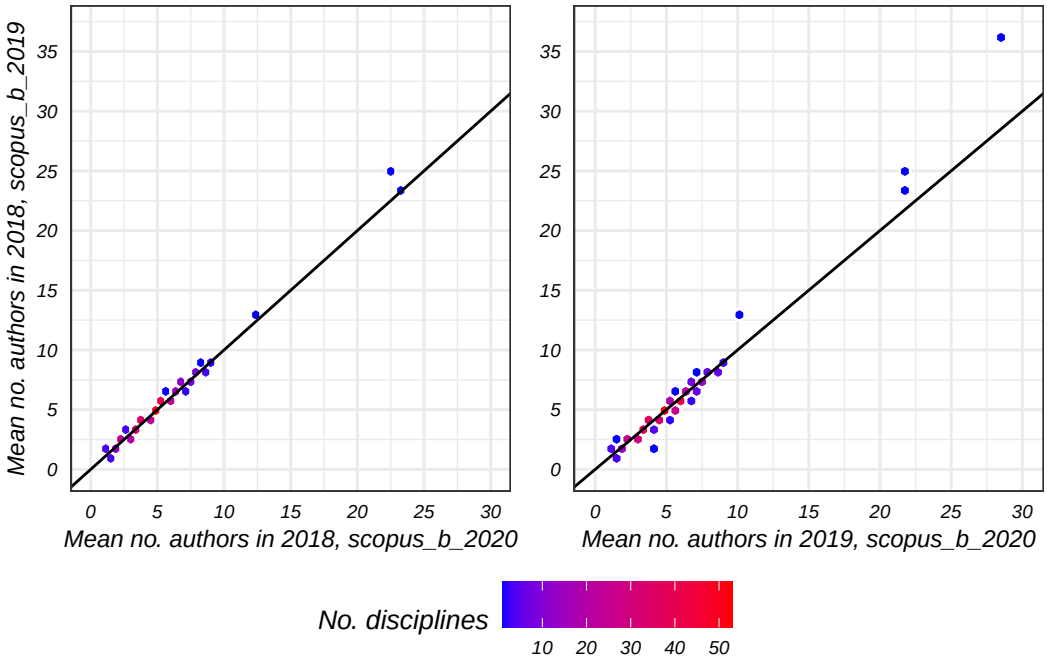

Figure 15: Mean number of authors per discipline between databases, where colour denotes the number of disciplines with this combination of mean authors.

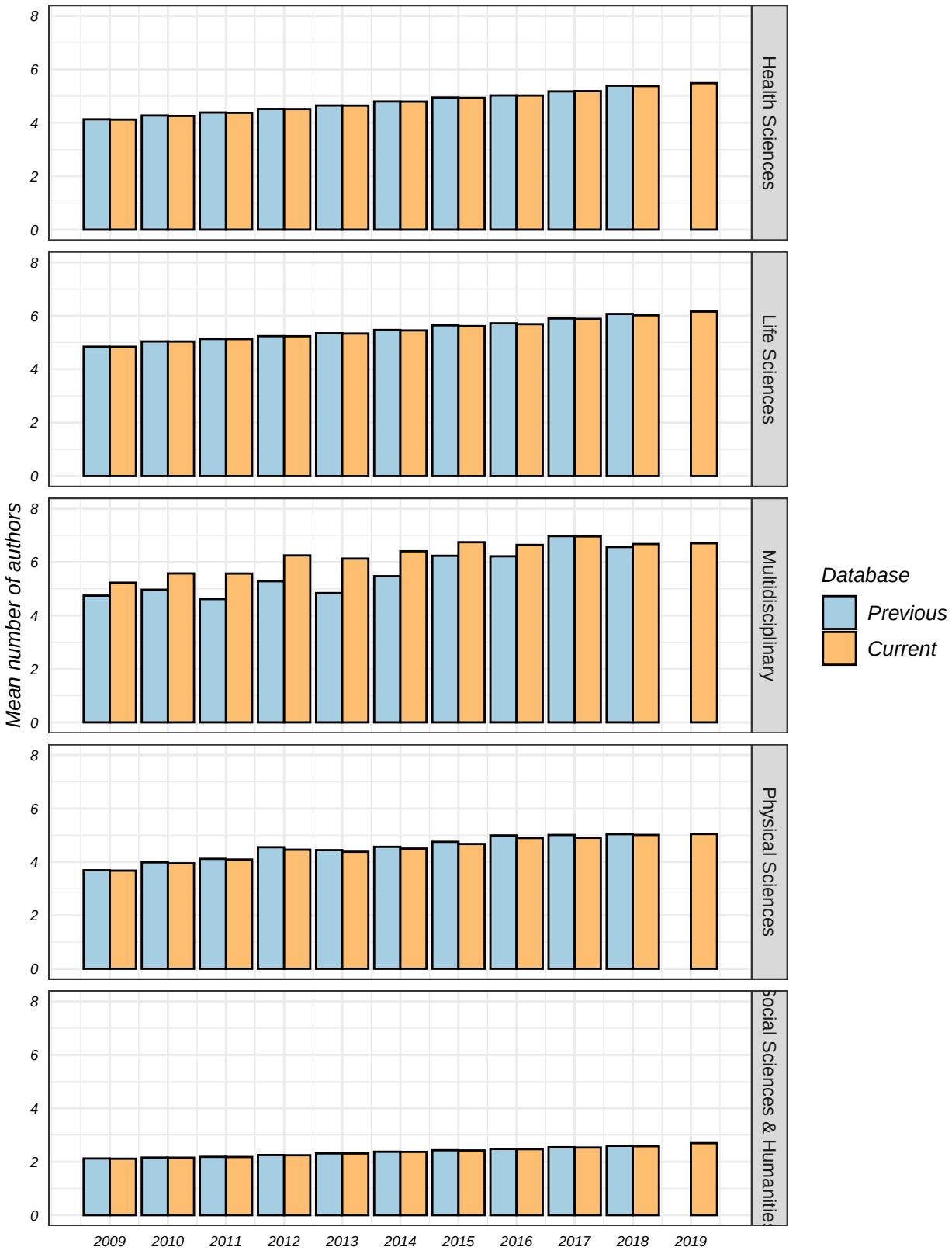

Figure 16: Annual mean number of authors by Subject Area and database.

Table 14: Disciplines where the mean number of authors changed by more than 10% between 2018 in scopus_b_2019 and 2019 in scopus_b_2020.

Discipline	Previous mean authors	Current mean authors	Perc. diff.	No. crrnt pubs.
Decision Sciences (miscellaneous)	2.2	4.1	84.3	210
Critical Care Nursing	3.3	3.9	17.5	1038
Economics and Econometrics	2.4	2.8	15.9	33892
Chemical Engineering (miscellaneous)	4.3	5.0	15.4	5044
Review and Exam Preparation	3.0	3.5	15.3	297
Materials Science (miscellaneous)	4.2	4.9	15.2	7430
Mathematical Physics	5.5	6.3	14.5	6143
Environmental Science (miscellaneous)	3.9	4.5	13.8	6077
LPN and LVN	3.3	3.7	13.8	1153
Health Professions (miscellaneous)	3.9	4.4	13.2	1409
Nuclear Energy and Engineering	5.1	5.7	12.8	11442
Research and Theory	3.2	3.6	12.8	460
General Arts and Humanities	2.0	2.3	11.3	5603
Ocean Engineering	3.9	4.3	10.5	9419
Economics, Econometrics and Finance (miscellaneous)	2.4	2.7	10.4	4638
Biotechnology	5.3	5.8	10.3	32965
General Health Professions	3.1	3.4	10.2	567
Chemical Health and Safety	5.4	4.9	-10.8	926
Engineering (miscellaneous)	25.3	21.5	-15.0	7285
Nuclear and High Energy Physics	35.8	28.7	-19.8	14299
Veterinary (miscellaneous)	12.6	10.0	-20.5	1279
Drug Guides	2.5	1.7	-30.2	47

Source items: Percentage by Subject 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 Elsevier 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 Scopus.

We show in the left-hand panel of Figure 17 the percentage of references that are source items per 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 scopus_b_2020 database compared to 2019 in the scopus_b_2020 database.

It is in this 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 15. Longer term trends can be seen in Figure 18 where we present the percentage of reference that are source items per Subject Area over the last ten common years of both databases.

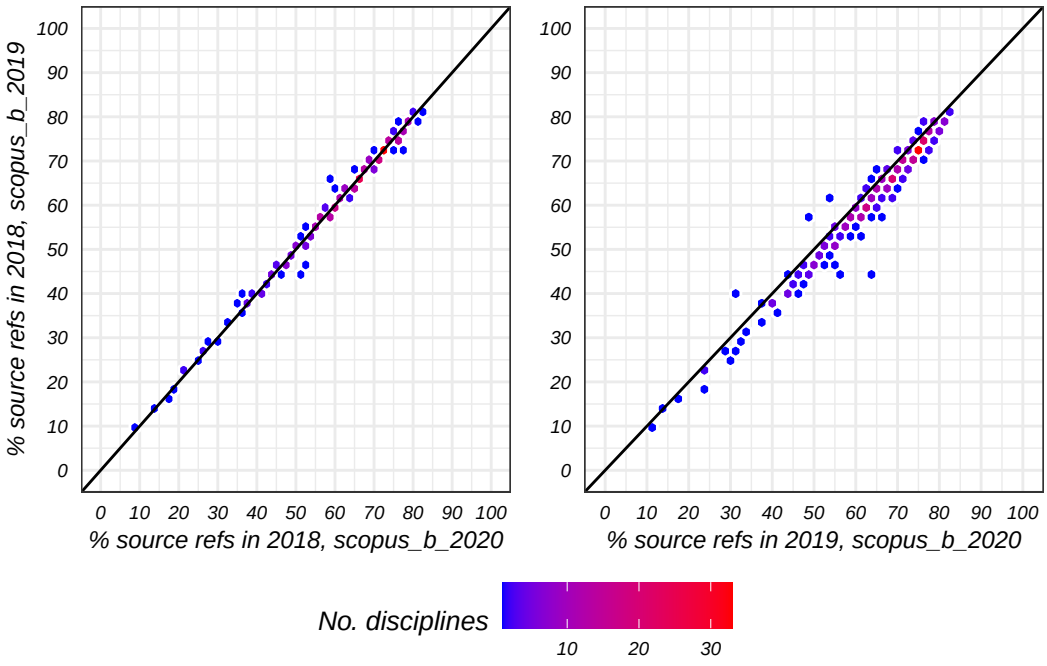

Figure 17: The percentage of cited items that are source items per discipline by database, where colour denotes the number of disciplines with this combination of source references.

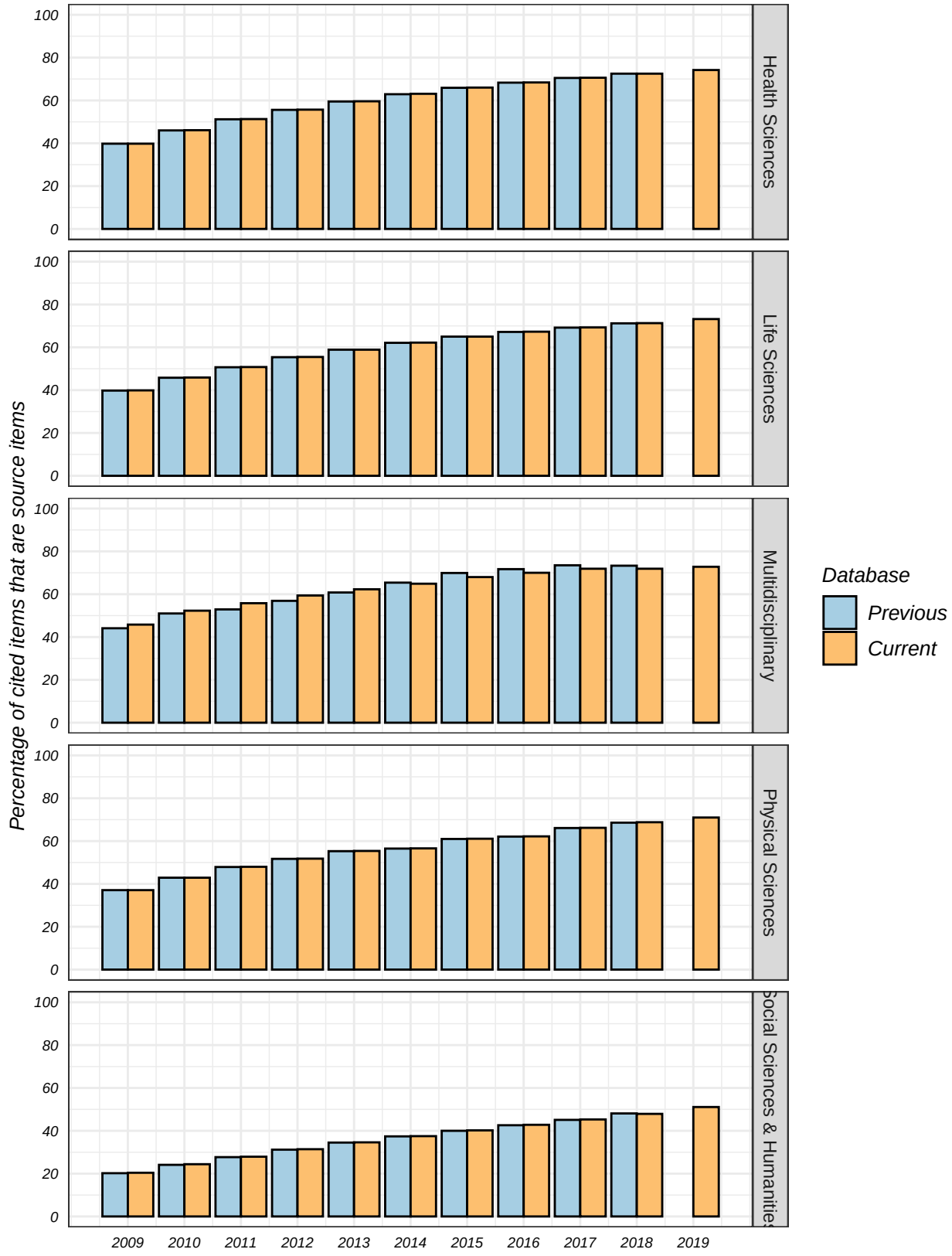

Figure 18: The annual percentage of cited items that are source items by Subject Area and database.

Table 15: Disciplines where the percentage of indexed references changed by more than 5 percentage points between 2018 in scopus_b_2019 and 2019 in scopus_b_2020.

Discipline	Previous no. refs.	Current no. refs.	Prvs % source	Crrnt % source	Change
Decision Sciences (miscellaneous)	9,760	40,165	44.8	63.5	18.7
Veterinary (miscellaneous)	763,201	564,759	43.8	57.4	13.6
Engineering (miscellaneous)	9,889,775	11,606,787	56.6	66.1	9.5
Computer Science (miscellaneous)	928,035	710,134	53.1	61.9	8.8
Economics and Econometrics	3,485,207	5,434,141	47.3	55.9	8.6
Hardware and Architecture	2,487,527	2,847,563	62.1	69.2	7.1
Nurse Assisting	9,129	8,339	58.4	64.7	6.3
Dental Hygiene	12,775	9,510	57.3	63.4	6.1
Architecture	387,756	595,186	42.4	48.4	6.0
Surfaces and Interfaces	2,563,014	4,475,513	71.3	76.9	5.6
Geometry and Topology	368,226	470,962	48.6	54.1	5.5
Physics and Astronomy (miscellaneous)	23,343,752	24,622,941	62.8	68.3	5.5
Food Science	9,280,168	11,453,654	64.2	69.4	5.2
Chemical Engineering (miscellaneous)	738,865	1,392,460	73.2	78.4	5.2
Management Science and Operations Research	1,460,109	1,878,203	53.5	58.7	5.2
Mathematical Physics	1,307,129	1,634,876	61.7	54.8	-6.9
Medical Assisting and Transcription	2,118	806	57.8	49.1	-8.7
Medical Terminology	588	1,062	40.0	30.8	-9.2

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.