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

Dimity Stephen, Stephan Stahlschmidt and Paul Donner

September 2021

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 2021

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 in key variables between the latest two iterations of the databases at the macro-level of institutions, sectors, countries, disciplines.

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, so that these factors may 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 examination 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_2020 and scopus_b_2021 databases is 2019, 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 2017 for the scopus_b_2020 and scopus_b_2021 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 2017 from the scopus_b_2020 database against 2018 in the scopus_b_2021 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 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 German institutions are also able to be examined via the KB's institutional coding for Germany. However, as there are a large number of institutions, we present data only for institutions that have shown substantial changes in the indicator of interest.

Methodological details

We focus on articles and reviews published in journals as these are the most common documents used in bibliometric analyses. Unless otherwise stated, we examine content indexed in the Science Citaton Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (A&HCI) WoS indices. 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.

Data for disciplines are presented based on either the All Science Journal Classification (ASJC) or the Subject Area classification. The ASJC is the fine-grained 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. Consequently, blank tables may appear in this report, but they are nonetheless informative about the indicator under examination.

Analysis

Publication counts: Total, selected countries, German sectors, and Subject Areas

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 Figure 1 we show the total number of documents of different types indexed in each database, followed by the whole counts of articles and reviews published by selected countries and German sectors over the last 10 years in Figures 2 and 3. In Figure 4 we show the distribution of publications by Subject Area.

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_2021 database contains a greater number of publications for the most recent years than the scopus_b_2020 database due to the continued indexing of items by Elsevier past the annual point in April at which the 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. 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 region-, country-, or discipline-specific journals. Decreases may reflect the de-indexation of journals 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 or decreases in the current database in recent years may warrant investigation.

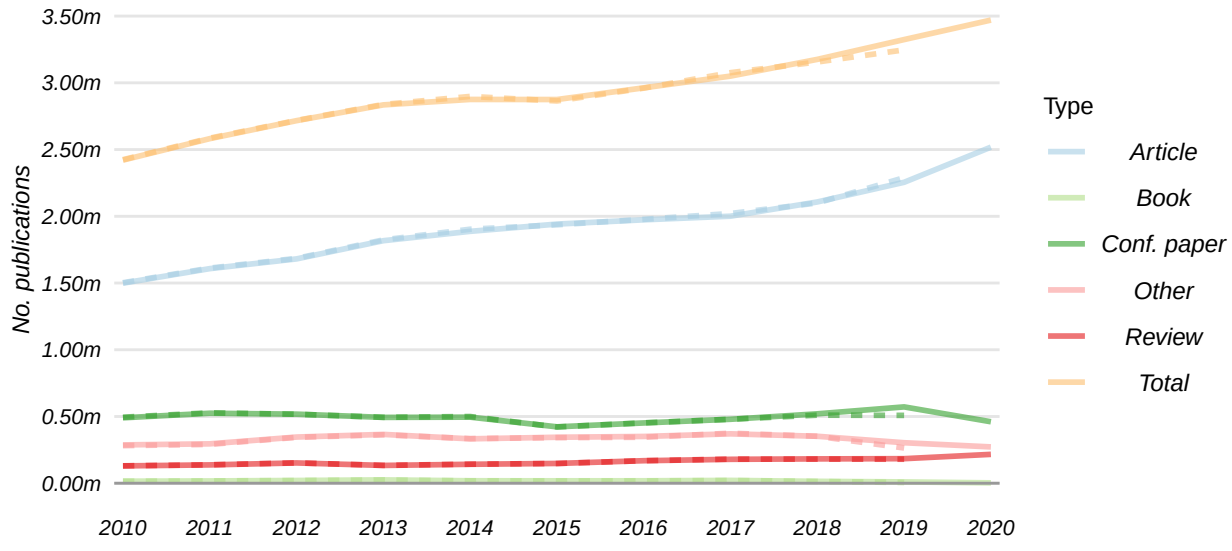

Figure 1: Number of publications by document type and database, where dashed lines show the previous database and full lines show the current database.

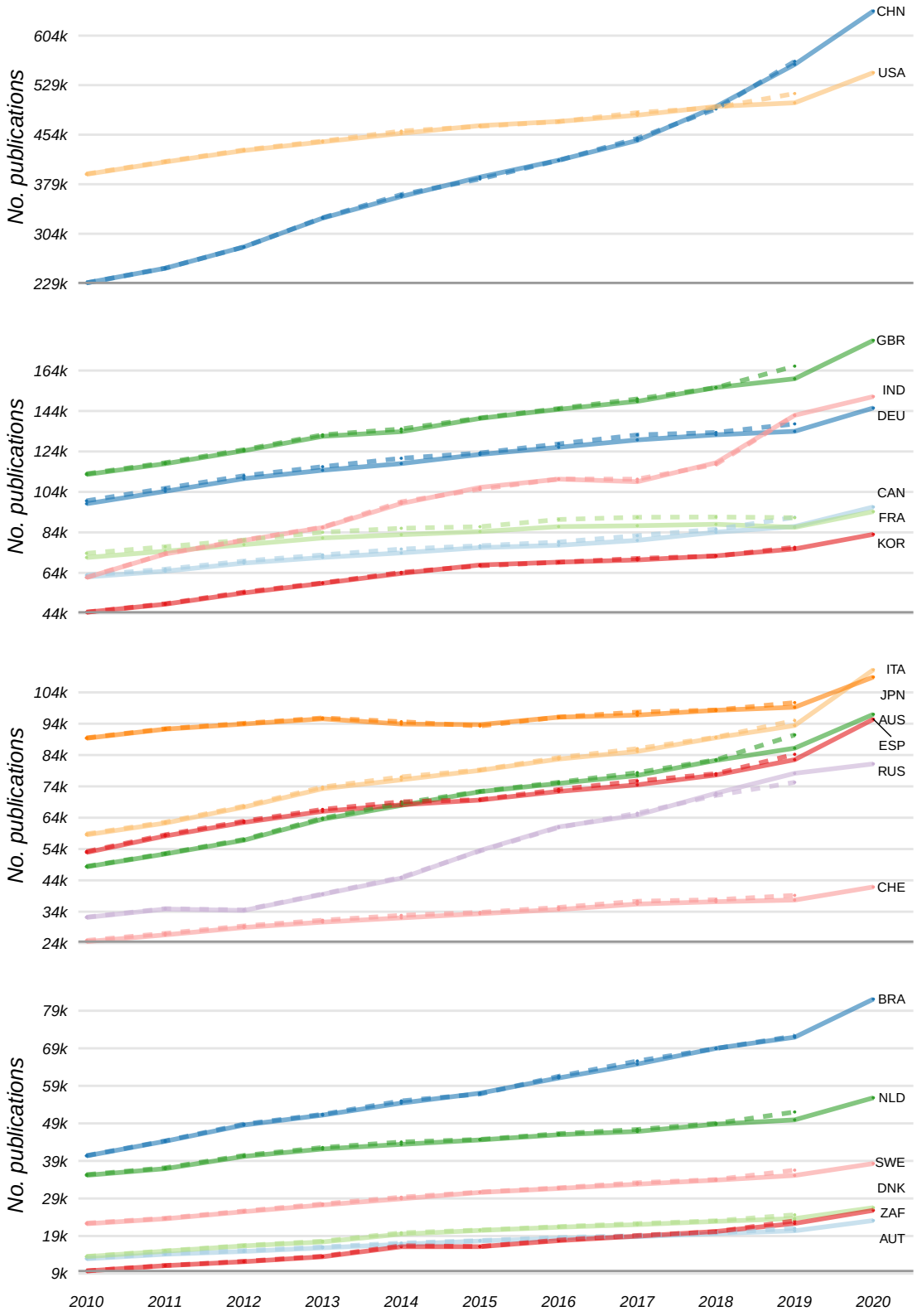

Figure 2: Whole counts of national publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels have different axes.

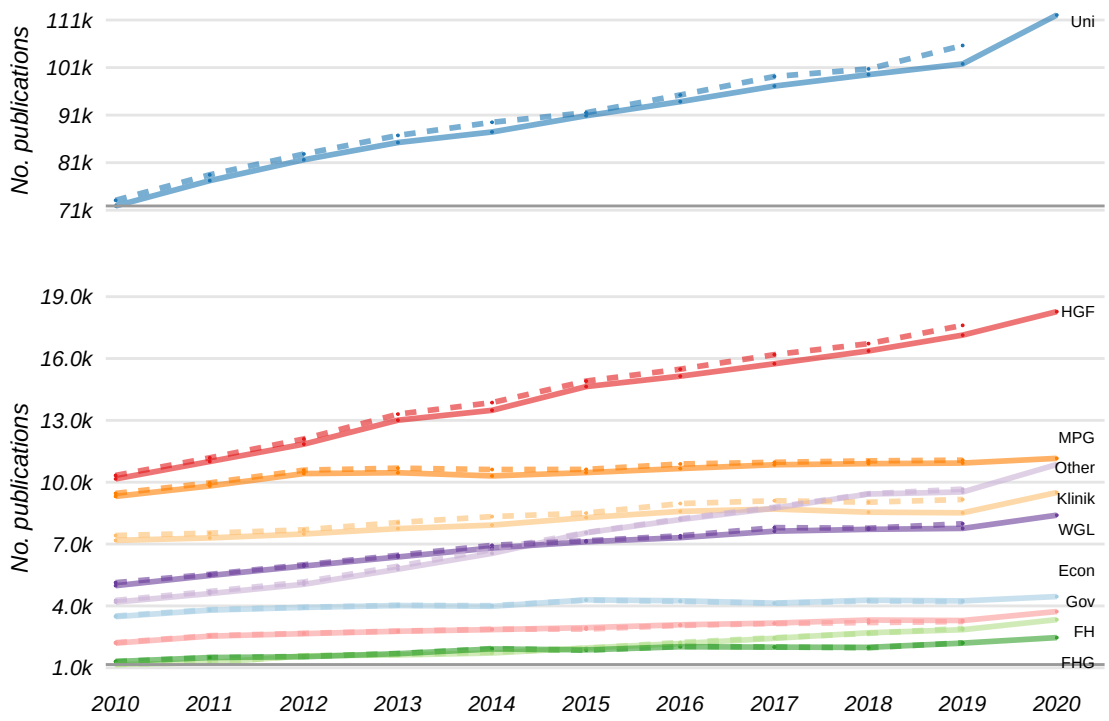

Figure 3: Whole counts of sectoral publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels' different scales.

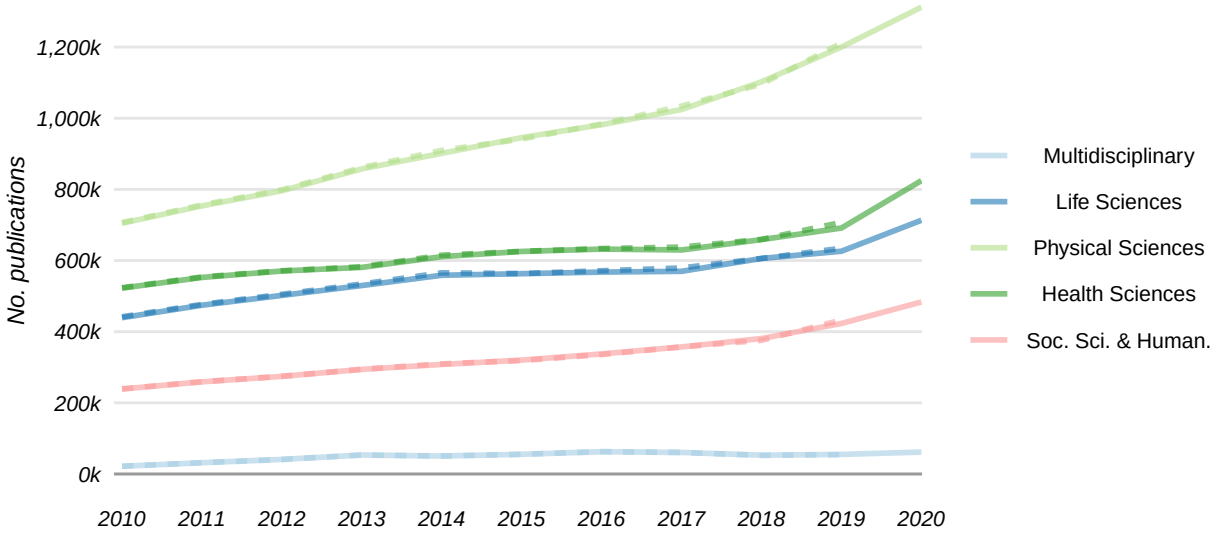

Figure 4: Whole counts of publications by Subject Area and database, where dashed lines show the previous database and full lines show the current database.

Journals: Total indexed and numbers added or removed

The journals indexed constitute the foundation of the database. Year to year changes in the journals indexed reflect the database provider's curation procedures to introduce new content and remove content no longer meeting indexation criteria. The amount of and changes in content indexed can influence bibliometric indicators, such as country-level counts of publications and citations, particularly if changes are concentrated in specific disciplines.

As all sources indexed have titles – as opposed to some missingness of ISSNs – changes in journals were identified by matching the titles of all journals indexed in 2019 in the scopus_b_2020 database to those with 2020 content in the scopus_b_2021 database. Titles in scopus_b_2020 but not scopus_b_2021 were considered removed, while titles present in scopus_b_2021 but not in scopus_b_2020 were considered added. These data may include a small number of journals that changed titles. Journals were mapped to the Subject Areas classification from the ASJC classification. Some double-counting of journals between SAs occurs where the journal is assigned to two or more classifications mapped to different SAs. In total, 2542 journals were added and 1403 were removed.

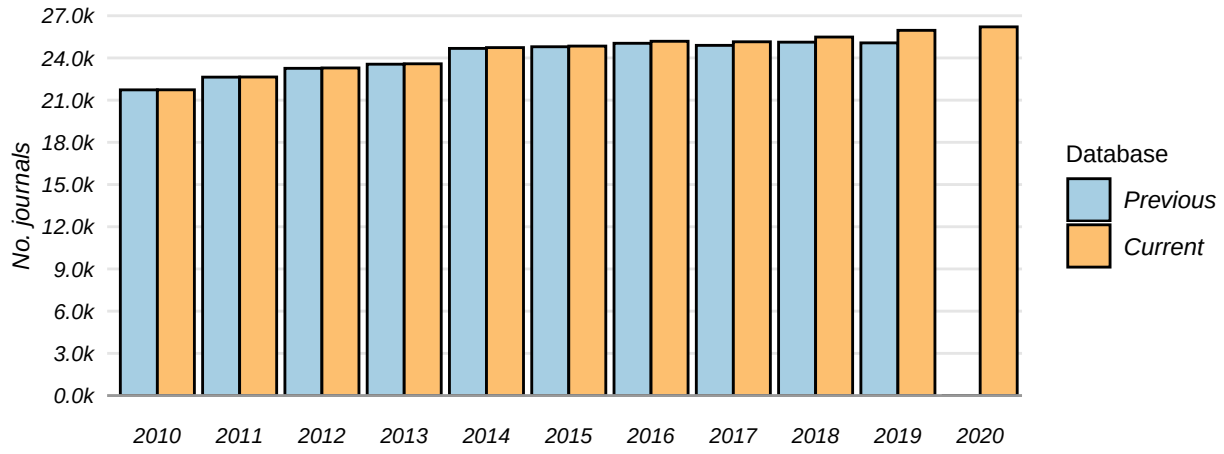

Figure 5: The number of journals indexed in each source over time.

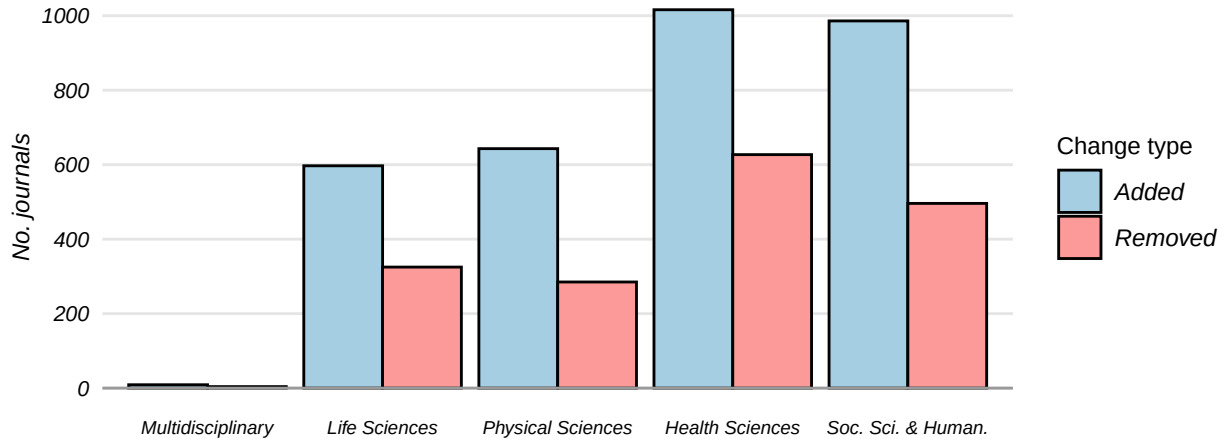

Figure 6: The number of journals added and removed between the latest years in each database by Subject Area.

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 for the most recent years from the two databases are presented for German sectors in Figure 7 and for countries in Figure 8. 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.

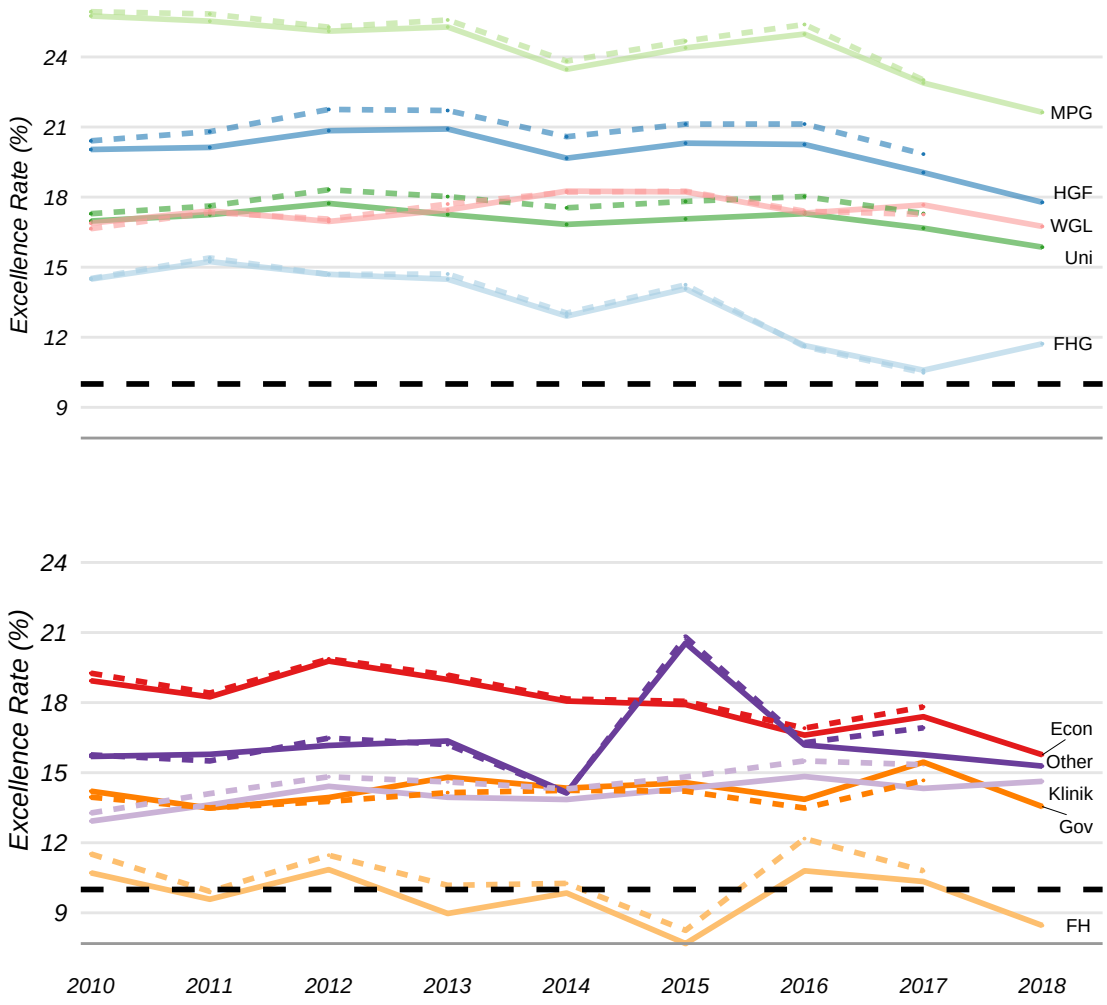

Figure 7: Excellence rates by sector, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

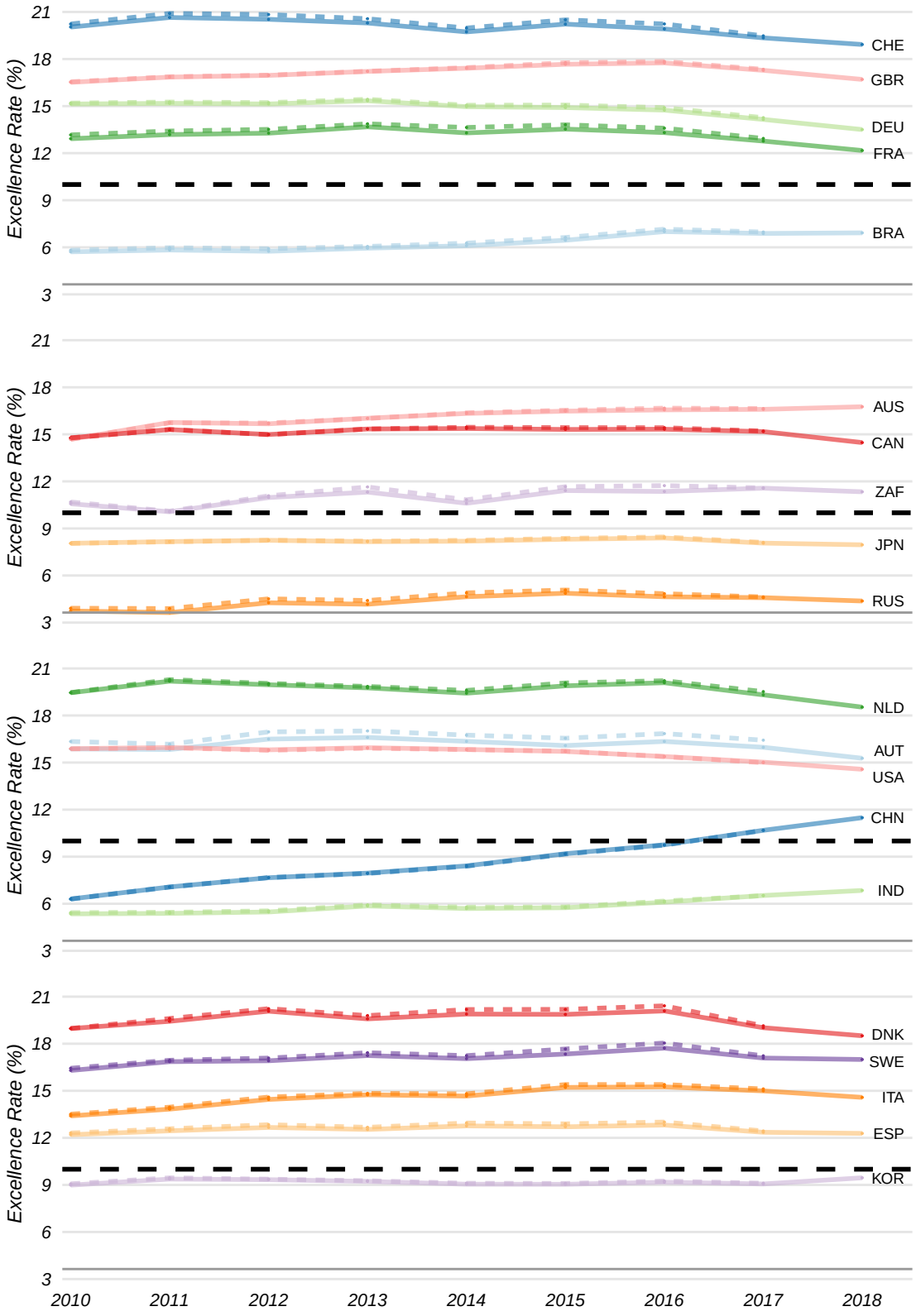

Figure 8: Excellence rates for selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

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]. This results from 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 greater consistency of coverage in Scopus, 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 9 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 9 we see the ER thresholds for each discipline in 2017 in both the scopus_b_2020 and scopus_b_2021 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 2017 in Figures 7 and 8. In the bottom panels we present again the thresholds for each discipline in 2017 in the scopus_b_2020 database but now compared against the threshold in 2018 in the scopus_b_2021 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 9 are shown in Tables 1 and 2, along with disciplines where the previous threshold was zero, highlighting potentially new or emerging disciplines.

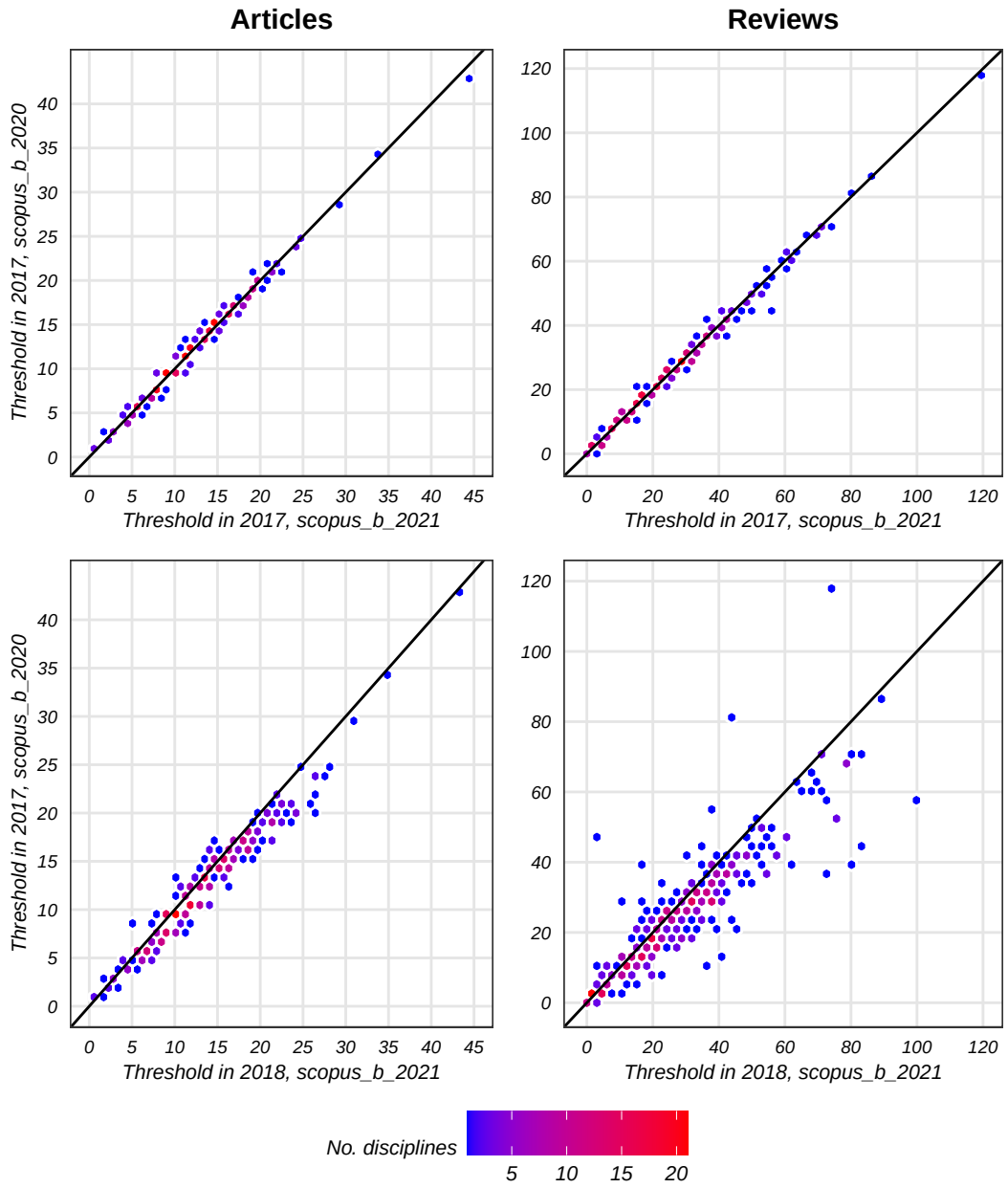

Figure 9: The ER threshold for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of thresholds.

Table 1: Articles: Disciplines where the ER threshold decreased, or increased by over 40% between 2017 in scopus_b_2020 and 2018 in scopus_b_2021, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Medical Terminology	1	2	47	100.0
Nurse Assisting	2	3	139	50.0
Review and Exam Preparation	4	6	239	50.0
Critical Care and Intensive Care Medicine	16	15	6,109	-6.2
Hardware and Architecture	15	14	18,961	-6.7
Earth-Surface Processes	14	13	10,713	-7.1
Behavioral Neuroscience	13	12	5,875	-7.7
Advanced and Specialized Nursing	13	12	2,995	-7.7
Anatomy	12	11	3,598	-8.3
Neuropsychology and Physiological Psychology	12	11	3,472	-8.3
Accounting	11	10	4,831	-9.1
Chemical Health and Safety	10	9	724	-10.0
Logic	9	8	1,159	-11.1
Rheumatology	17	15	4,885	-11.8
Nephrology	16	14	4,491	-12.5
Transplantation	16	14	4,367	-12.5
Community and Home Care	5	4	1,294	-20.0
Emergency Medical Services	5	4	160	-20.0
Computer Science (miscellaneous)	9	7	8,926	-22.2
Paleontology	13	10	3,933	-23.1
Drug Guides	4	3	85	-25.0
Medical Assisting and Transcription	3	2	44	-33.3
Pharmacology (nursing)	8	5	255	-37.5

Table 2: Reviews: Disciplines with a current or previous ER threshold of at least 5, where the threshold decreased by over 25%, increased by over 60% between 2017 in scopus_b_2020 and 2018 in scopus_b_2021, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Medical Assisting and Transcription	0	2	4	Inf
Numerical Analysis	12	40	35	233.3
Health Professions (miscellaneous)	11	36	68	227.3

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
General Health Professions	7	22	6	214.3
Decision Sciences (miscellaneous)	6	16	7	166.7
Emergency Medical Services	4	10	13	150.0
Assessment and Diagnosis	8	19	37	137.5
Care Planning	8	19	36	137.5
Transportation	21	46	174	119.0
Signal Processing	39	79	201	102.6
Theoretical Computer Science	36	72	202	100.0
Nurse Assisting	3	6	19	100.0
Complementary and Manual Therapy	6	12	59	100.0
Materials Science (miscellaneous)	23	44	317	91.3
Economic Geology	20	38	75	90.0
Review and Exam Preparation	8	15	5	87.5
Nuclear Energy and Engineering	45	82	295	82.2
Food Animals	15	27	178	80.0
Media Technology	4	7	111	75.0
Equine	8	14	97	75.0
General Mathematics	15	26	270	73.3
Geotechnical Engineering and Engineering Geology	18	31	372	72.2
Leadership and Management	7	12	107	71.4
Computational Mechanics	17	29	56	70.6
Pharmacy	10	17	106	70.0
Ecological Modeling	59	100	106	69.5
Environmental Science (miscellaneous)	19	31	350	63.2
Information Systems	24	39	552	62.5
Safety, Risk, Reliability and Quality	22	35	375	59.1
Computer Science (miscellaneous)	16	25	108	56.2
Horticulture	18	28	251	55.6
Fundamentals and Skills	11	17	58	54.5
Safety Research	19	29	322	52.6
General Computer Science	40	61	593	52.5
Health Information Management	21	32	111	52.4
Computer Vision and Pattern Recognition	41	30	95	-26.8
Agricultural and Biological Sciences (miscellaneous)	22	16	306	-27.3
Industrial Relations	7	5	145	-28.6
Multidisciplinary	55	39	1,521	-29.1
Mathematics (miscellaneous)	10	7	57	-30.0
Statistics and Probability	23	16	264	-30.4
Computer Graphics and Computer-Aided Design	34	23	181	-32.4
Research and Theory	9	6	53	-33.3

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Earth and Planetary Sciences (miscellaneous)	26	17	155	-34.6
Energy (miscellaneous)	119	75	270	-37.0
Mathematical Physics	30	17	161	-43.3
Biochemistry, Genetics and Molecular Biology (miscellaneous)	81	45	321	-44.4
Statistical and Nonlinear Physics	38	17	99	-55.3
Geometry and Topology	28	11	24	-60.7
Psychology (miscellaneous)	11	2	67	-81.8
Discrete Mathematics and Combinatorics	46	4	14	-91.3

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

The number of citations a publication could be expected to receive is dependent to an extent 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 accrued in the first 3 years after publication. As we did with ERs, we examine here in Figure 10 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 10 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 levels of citations.

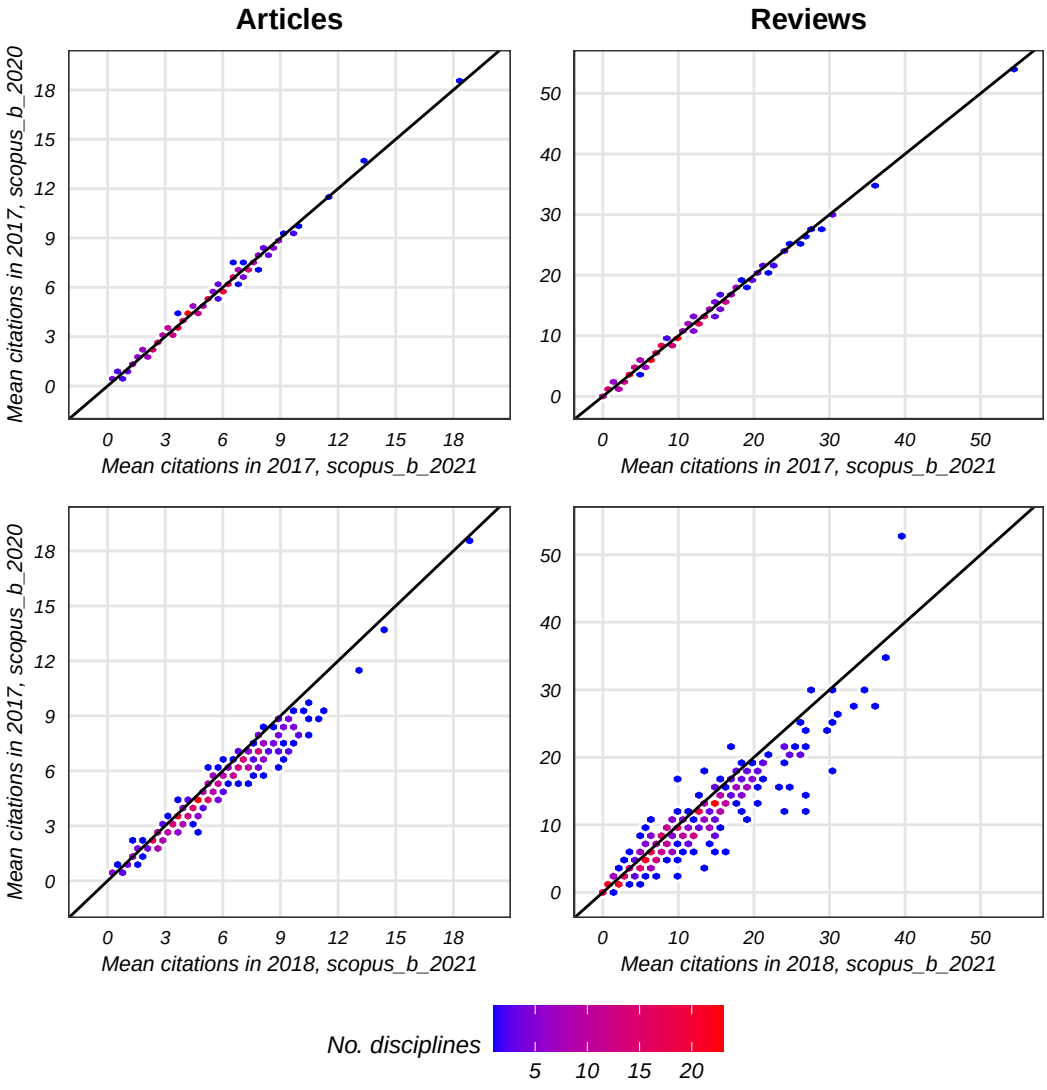

Figure 10: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

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 2017 in scopus_b_2020 and 2018 in scopus_b_2021, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Nurse Assisting	0.7	1.7	139	150.01
Histology	2.7	4.9	4,367	80.23
Information Systems	5.6	8.2	16,789	47.58
Research & Theory	2.1	3.0	369	46.37
Review & Exam Preparation	1.8	2.5	239	44.98
Chemistry (misc.)	6.2	8.7	5,374	40.61
Dental Hygiene	1.9	2.6	82	37.64
Leadership & Mgmnt	1.9	2.6	1,461	37.02
General Business, Mgmnt & Accounting	2.7	3.7	8,151	36.82
Hematology	7.0	9.6	10,721	36.36
General Arts & Humanities	0.7	1.0	4,407	36.20
Dentistry (misc.)	2.7	3.6	696	35.87
General Engineering	3.2	4.3	50,181	35.34
Waste Mgmnt & Disposal	7.0	9.4	18,389	34.71
Health Information Mgmnt	6.8	9.2	1,702	34.17
Periodontics	5.5	7.2	1,000	32.65
General Energy	8.0	10.4	12,041	30.66

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 2017 in scopus_b_2020 and 2018 in scopus_b_2021, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Medical Assisting & Transcription	0.0	0.8	4	Inf
Health Professions (misc.)	3.3	13.1	68	294.26
General Health Professions	2.4	9.4	6	283.03
Media Technology	1.6	4.8	111	195.80
Emergency Medical Services	1.2	3.3	13	179.32
Review & Exam Preparation	2.6	7.0	5	170.97
Numerical Analysis	6.0	16.1	35	166.56
Transportation	5.8	15.2	174	163.39
Decision Sciences (misc.)	2.0	4.7	7	136.68
Assessment & Diagnosis	2.3	5.2	37	132.22
Care Planning	2.0	4.4	36	118.84
Theoretical Computer Science	12.3	26.4	202	114.36
Horticulture	6.1	11.9	251	95.78
Nuclear Energy & Engineering	12.3	24.0	295	95.53

Discipline	Previous cit.	Current cit.	No. crrnt pubs	Perc. diff.
Stratigraphy	7.5	14.2	54	90.91
Food Animals	5.4	10.1	178	87.40
Building & Construction	10.1	18.9	586	86.16
General Energy	14.8	27.4	449	85.88
Chiropractics	3.9	6.9	26	76.46
General Mathematics	4.0	7.0	270	75.65
Earth-Surface Processes	4.2	7.3	293	74.13
General Decision Sciences	8.3	14.5	66	74.09
Economic Geology	8.4	14.6	75	72.94
Ecological Modeling	17.6	30.3	106	72.39
Dentistry (misc.)	4.5	7.8	98	72.34
Complementary & Manual Therapy	2.0	3.3	59	68.40
Pharmacy	4.3	7.2	106	66.62
Computer Science (misc.)	7.0	11.2	108	60.95
Earth & Planetary Sciences (misc.)	8.6	6.8	155	-20.99
Veterinary (misc.)	11.1	8.8	145	-21.04
Nursing (misc.)	5.6	4.4	77	-21.73
Statistics, Probability & Uncertainty	11.2	8.7	126	-22.43
Computer Vision & Pattern Recognition	21.8	16.8	95	-22.78
Accounting	7.1	5.4	180	-24.87
Energy (misc.)	53.4	39.6	270	-25.90
Physiology (medical)	18.0	12.8	1,830	-28.96
Statistics & Probability	8.4	5.6	264	-34.11
Radiological & Ultrasound Technology	10.9	7.0	344	-35.46
Fundamentals & Skills	6.2	3.9	58	-37.60
Statistical & Nonlinear Physics	10.0	6.2	99	-37.94
Computer Graphics & Computer-Aided Design	16.3	9.8	181	-39.93

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 Analytics continues to index journals, the likelihood increases 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 11 and 12 the percentage of articles and reviews per German sector and selected country that remained uncited 3 years after they were published.

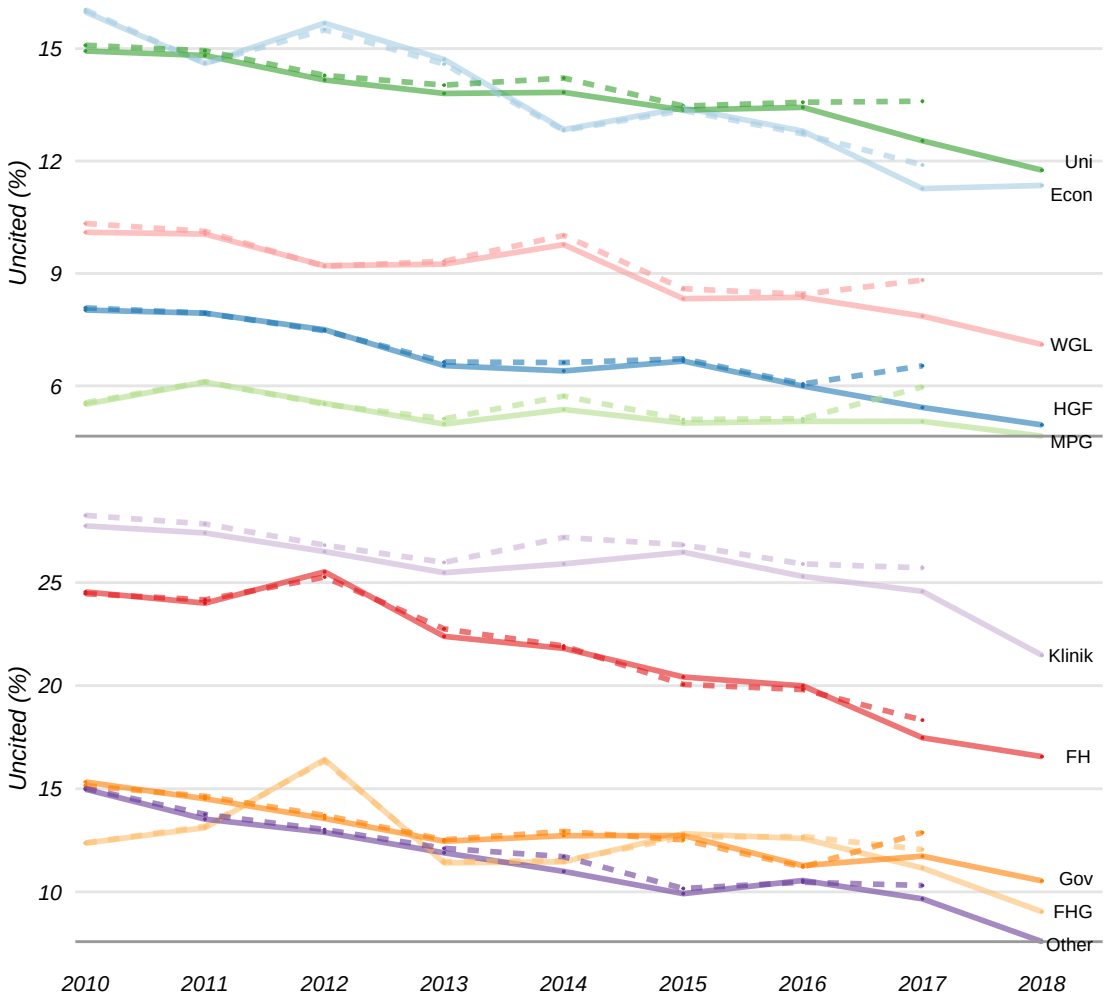

Figure 11: The percentage of uncited publications by German sector, based on whole counts, where dashed lines show the previous database and full lines show the current database.

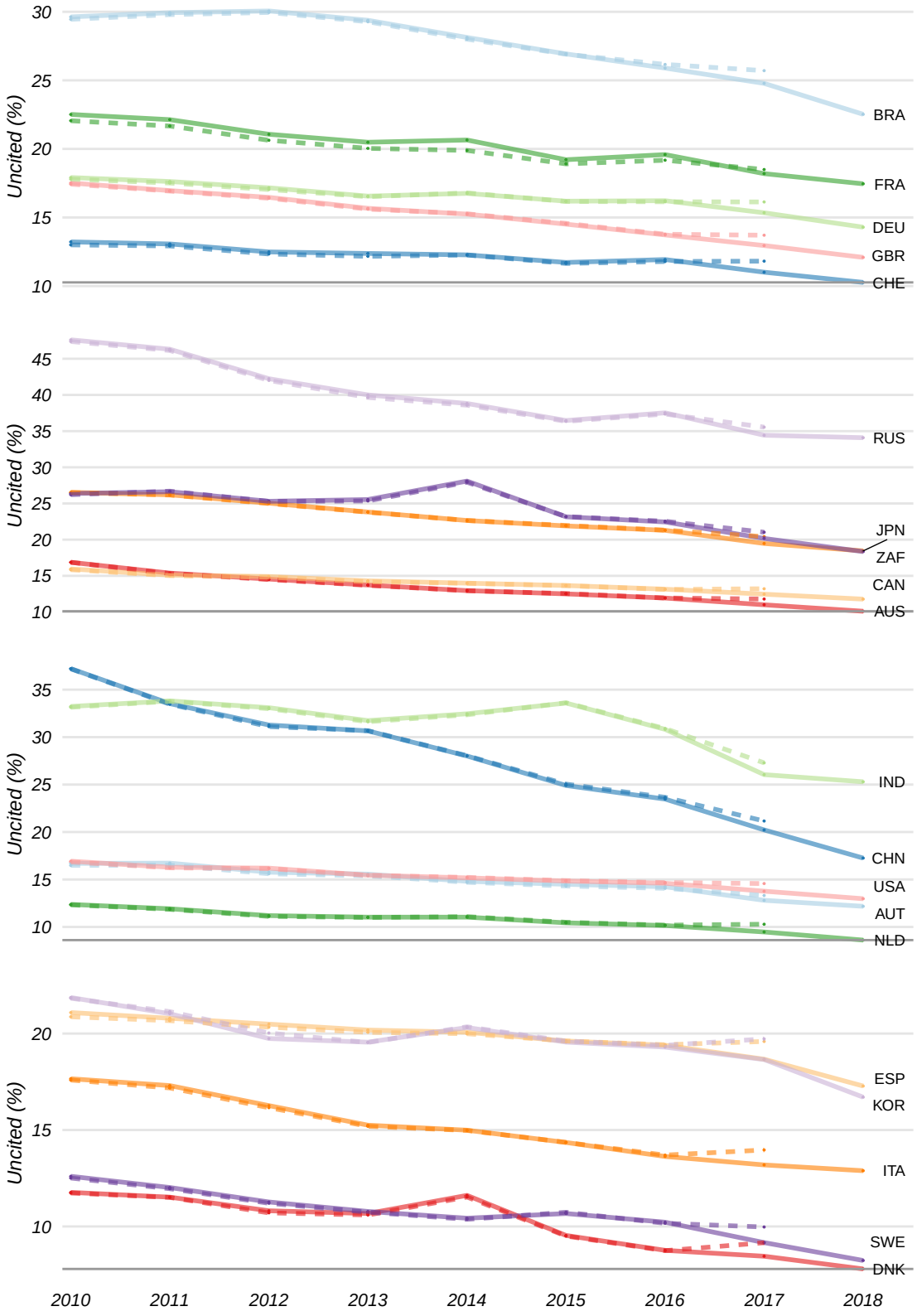

Figure 12: The percentage of uncited publications by selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database.

Disciplines: Changes in discipline classification

This section shows in Table 5 any changes that 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. 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 Elsevier's published classification structure. Figure 13 shows the number of publications assigned to specific disciplines identified to have changed in recent versions of the database.

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)	NA	29

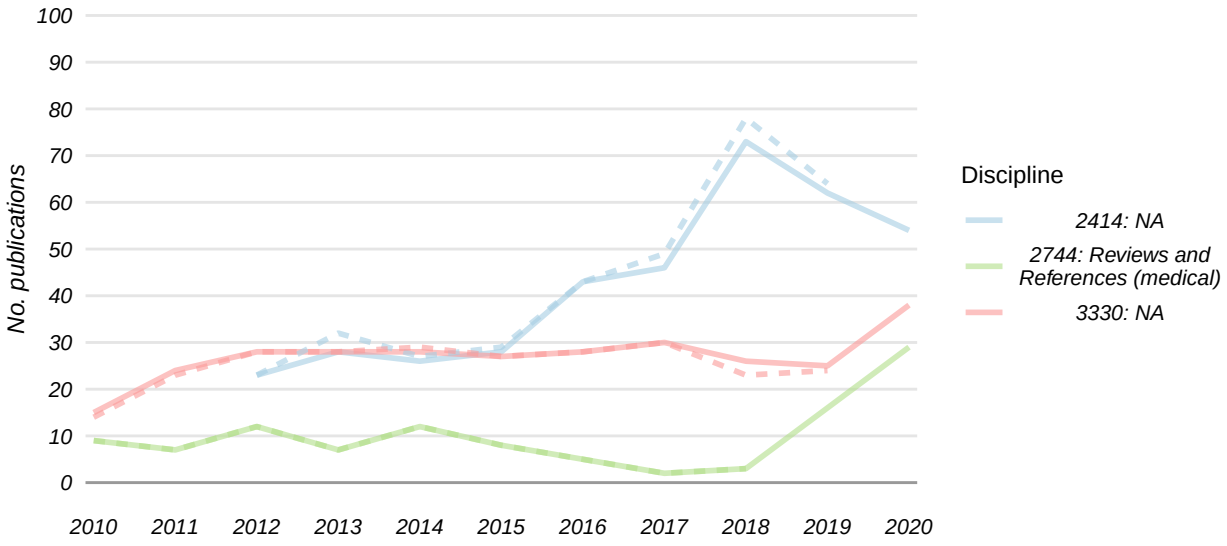

Figure 13: Time-series of disciplines known to have changed in recent versions of the databases. Dashed lines show the previous database and full lines show the current database.

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 14 the 20 disciplines with the highest percentage increases and decreases in publication counts between 2019 in scopus_b_2020 and 2020 in scopus_b_2021. 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.

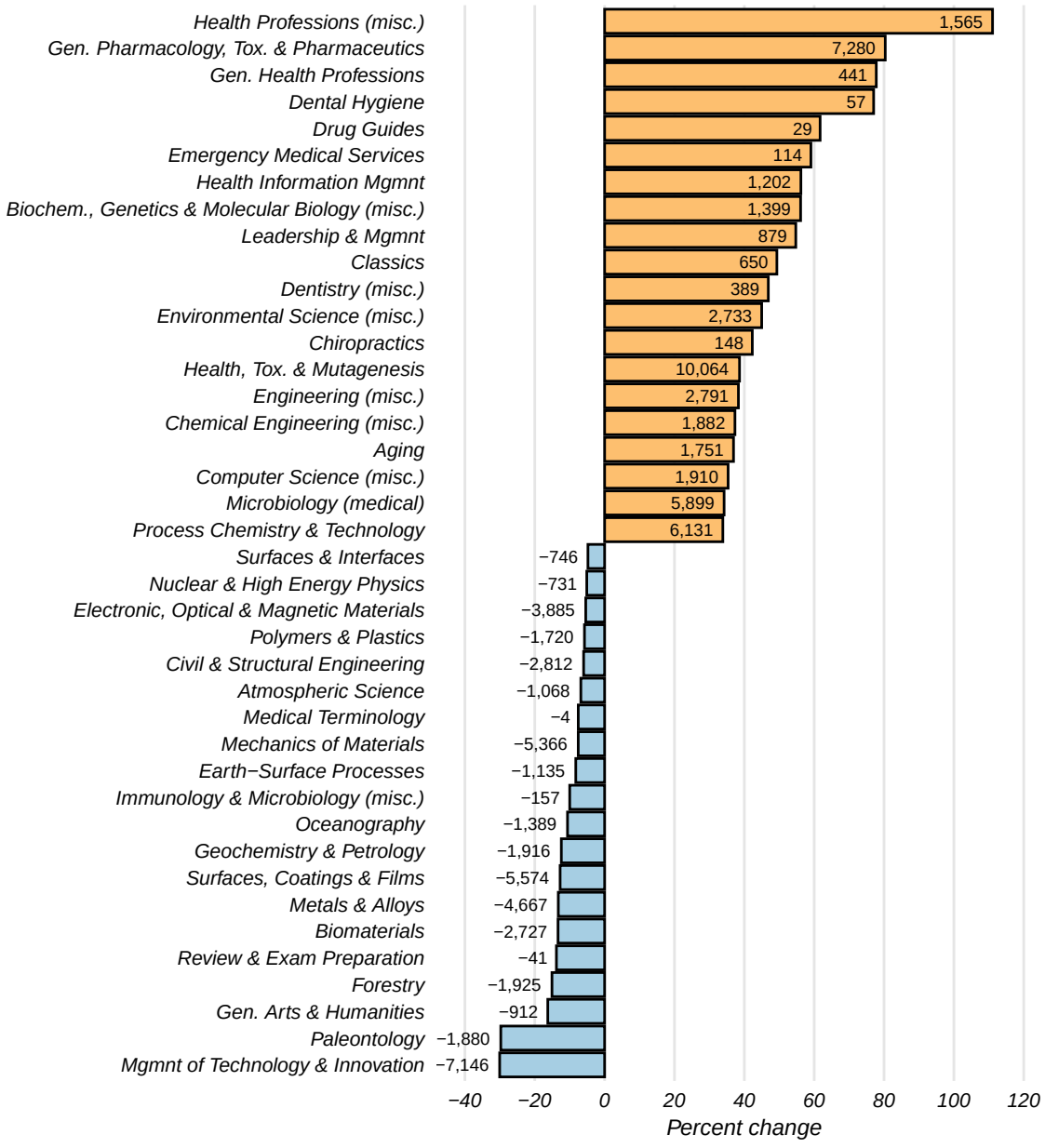

Figure 14: The 40 disciplines with the highest percentage change in publication counts between 2019 in scopus_b_2020 and 2020 in scopus_b_2021, with numerical difference in counts.

Disciplines: Number of publications not assigned to a discipline

This section presents in Figure 15 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.

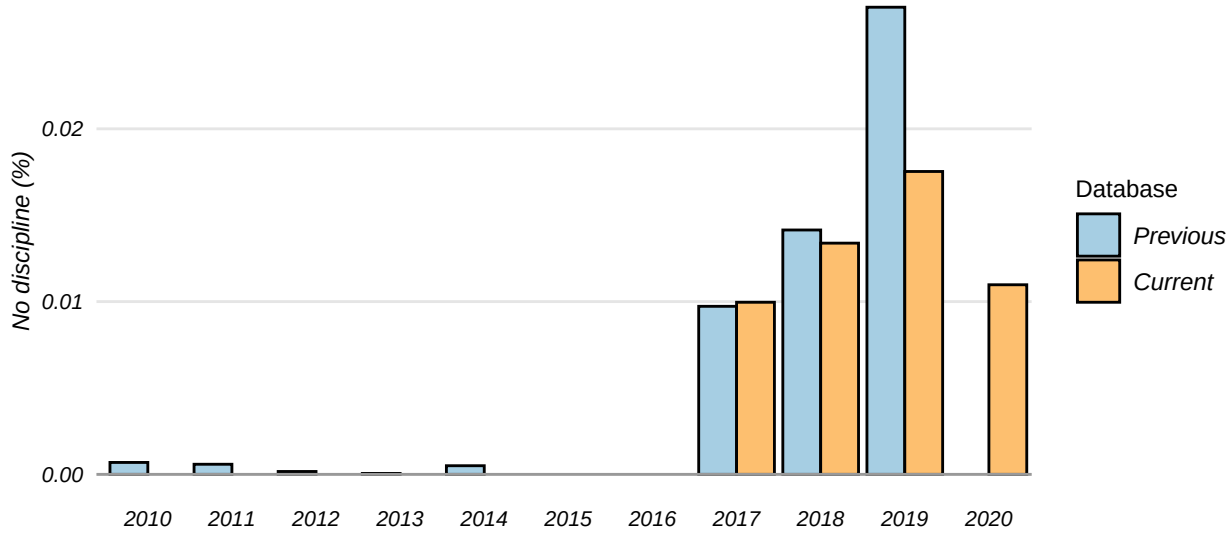

Figure 15: The percentage of publications in each database that do not have a discipline classification.

Metadata: Changes in pubyear, doctype, pubtype and items removed

This section details the number of items for which changes were made to key metadata in the latest iteration of the database or the items were removed. 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. We also examine the number of items that were present in the scopus_b_2020 database but not in the scopus_b_2021 database. 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 Analytics updates or corrects items, however changing or removing a large number of items may require investigation.

We identify changes in the metadata of in-scope items by first matching items between the scopus_b_2020 and scopus_b_2021 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. The number of items removed is based on UT_EIDs in scopus_b_2020 not matching a record in scopus_b_2021. Data are presented based on the publication year recorded in the

scopus_b_2020 database.

Table 6: The number of items with changes in metadata or removed between the previous and current database versions.

Year	Pub. year	Doc. type	Pub. type	Removed
2010	43	589	161	6,364
2011	70	622	42	7,550
2012	203	697	28	8,026
2013	1,736	436	23	8,517
2014	11,743	316	9	10,263
2015	1,628	489	232	11,789
2016	249	481	27	15,097
2017	293	509	30	36,323
2018	649	813	26	15,503
2019	90,366	5,514	107	106,395

Metadata: Missing metadata variables

Figure 16 shows 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. 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. Empty graphs indicate there were no items missing this metadata.

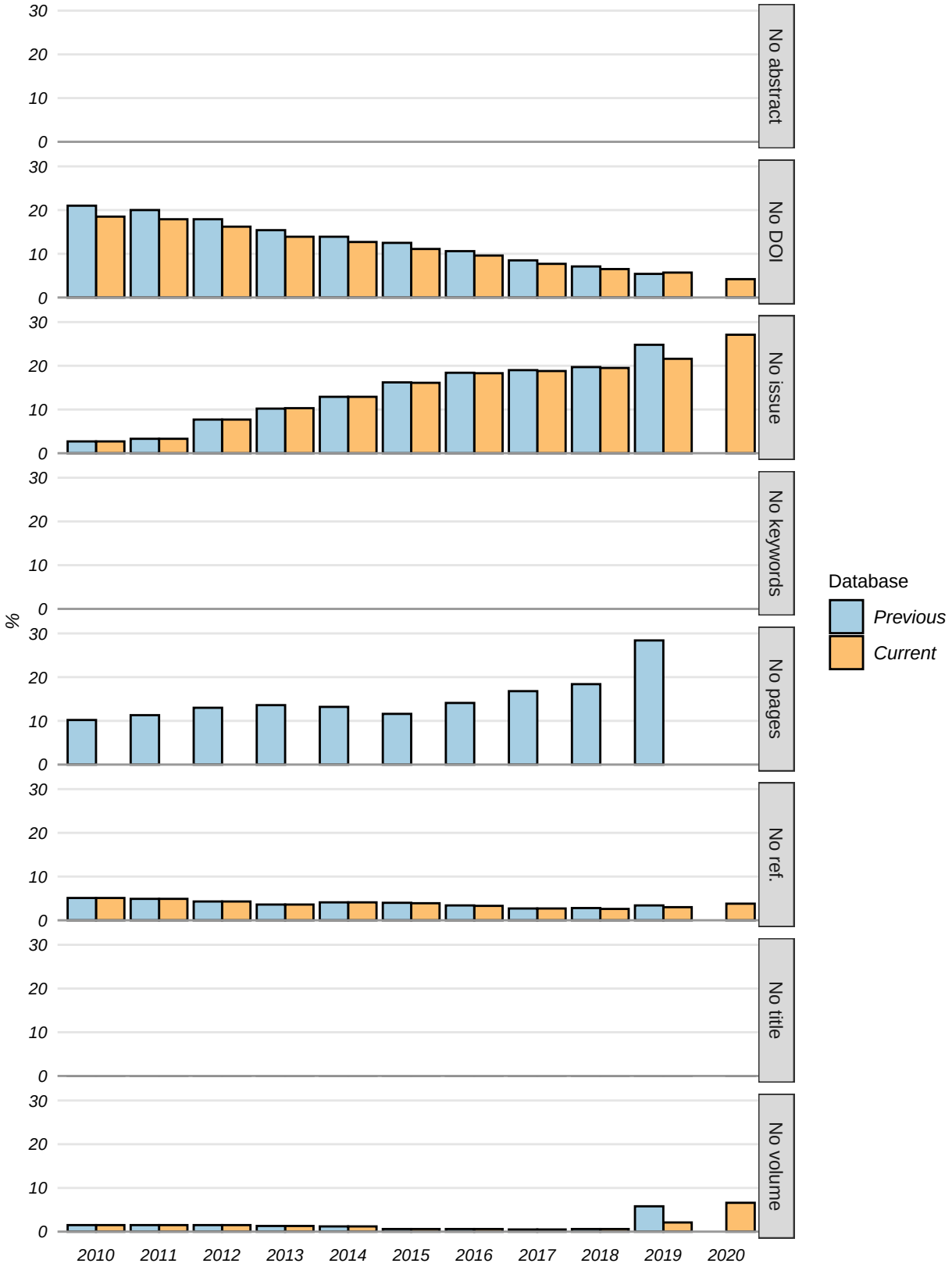

Figure 16: The percentage of items with missing metadata over time by database.

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. 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 KB 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 17 the number of items in each database between 2010 and 2019 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 17. Large disparities between the databases or substantial increases in missing information should be investigated.

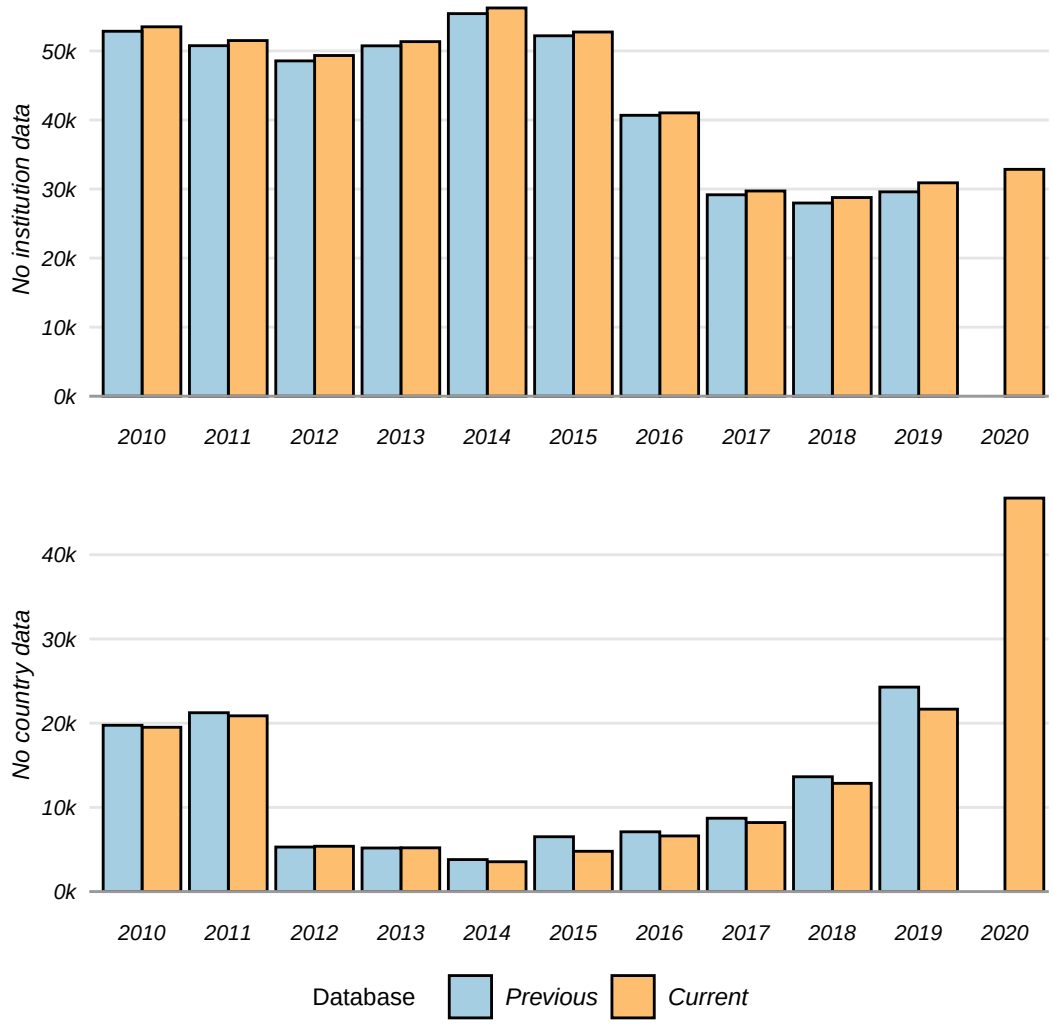

Figure 17: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

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 18, we see in the left panel the percentage of complete links for 2019 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 2019 in the scopus_b_2020, now compared with the 2020 in the scopus_b_2021, indicating potential changes between the latest year in each database.

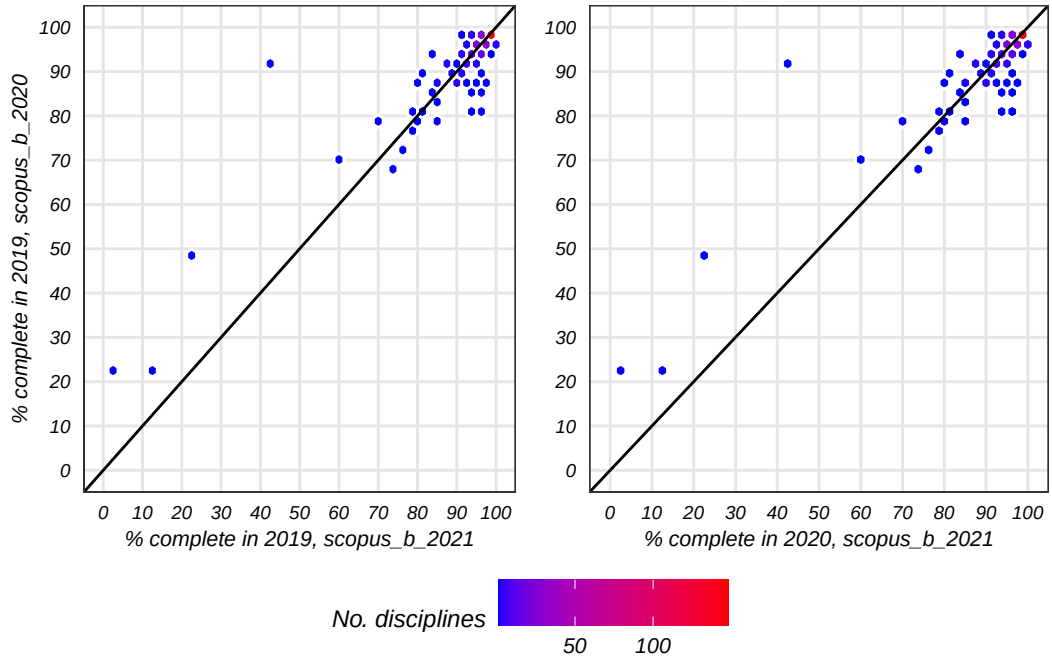

Figure 18: The percentage of complete author-institution links by disciplines.

The outlying disciplines observed in the right panel of Figure 18 that have changed by more than 7 percentage points in the percentage of complete author-institution links between databases are shown in Table 7.

Table 7: Disciplines (sc_extended) that changed by more than 7 percentage points in missing links between 2019 in scopus_b_2020 and 2020 in scopus_b_2021.

Discipline	Prvs items	% prvs complete	Crrnt items	% crrnt complete	Change
2806: Developmental Neuroscience	56,523	81.77	51,487	96.55	14.8
2911: Leadership and Management	22,057	81.35	121,958	94.97	13.6
3201: Psychology (miscellaneous)	25,119	85.53	30,656	95.49	10.0
2904: Care Planning	4,110	88.47	7,078	97.24	8.8
2720: Hematology	864,264	88.69	429,966	96.75	8.1
2914: Medical and Surgical Nursing	35,076	86.34	34,502	94.17	7.8
3205: Experimental and Cognitive Psychology	204,806	88.25	172,351	95.87	7.6
3602: Chiropractics	7,268	88.24	7,736	81.01	-7.2
3611: Pharmacy	62,276	88.74	68,968	81.20	-7.5
3401: Veterinary (miscellaneous)	32,538	70.11	40,251	60.90	-9.2
3503: Dental Hygiene	920	78.91	3,633	68.76	-10.1
1405: Management of Technology and Innovation	8,584,555	93.41	3,382,488	82.93	-10.5
2920: Pharmacology (nursing)	7,875	23.16	14,915	12.25	-10.9
3330: NA	148	22.97	256	2.34	-20.6
3606: Medical Assisting and Transcription	534	47.75	534	23.41	-24.3
2915: Nurse Assisting	2,216	91.11	2,887	41.67	-49.4

To provide context to the percentage of complete links observed in the most recent years, in Figure 19 we present the percentage of complete links made between authors and affiliations in each Subject Area over the last decade in both databases, plus 2020 in scopus_b_2021. Substantial changes between years or differences between the databases may require investigation of the cause.

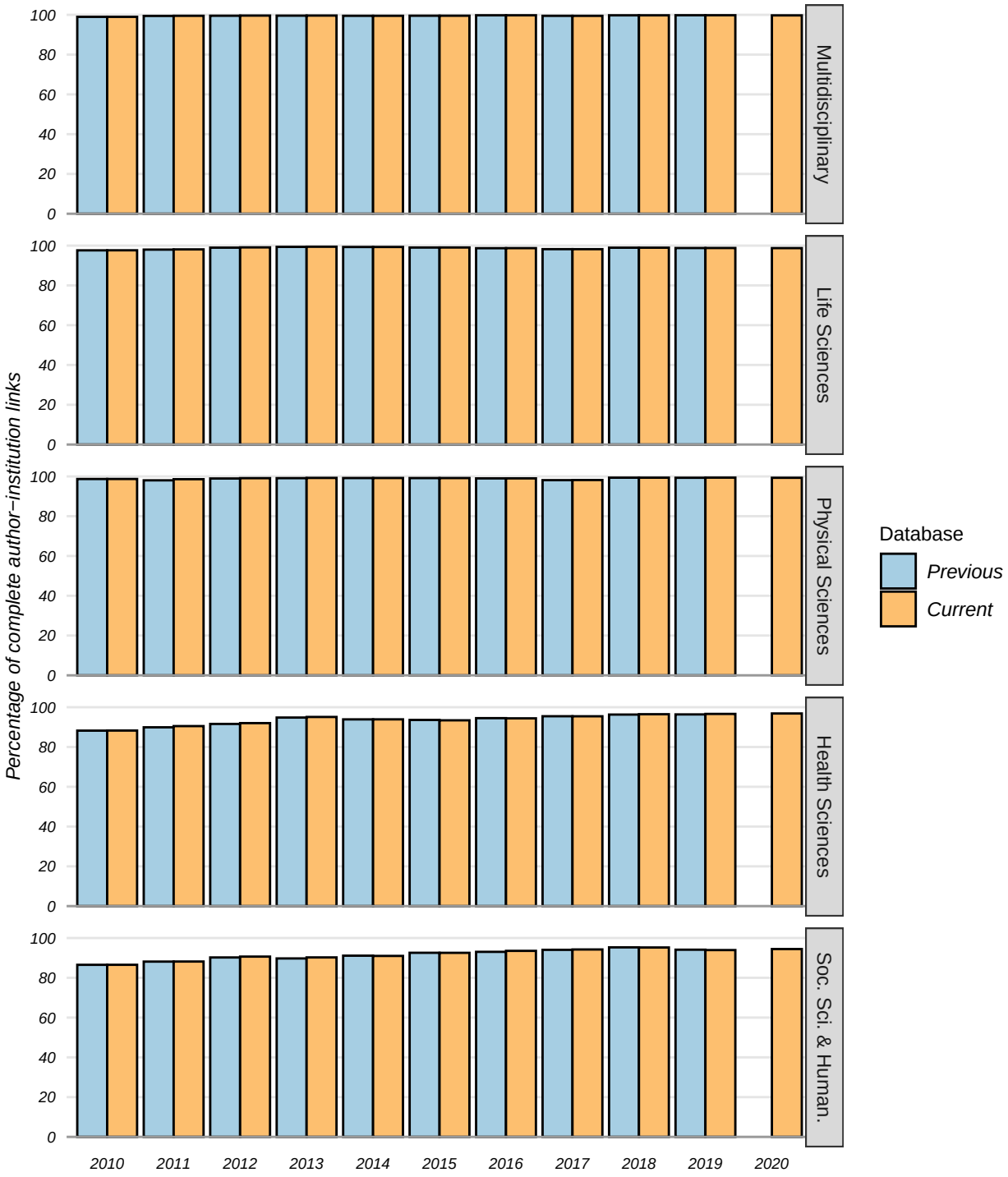

Figure 19: The annual percentage of complete author-institution links by Subject Area and database.

German institutions: German publications missing from KB institution coding

In Figure 20 we show the annual percentage 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 may be due to 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.

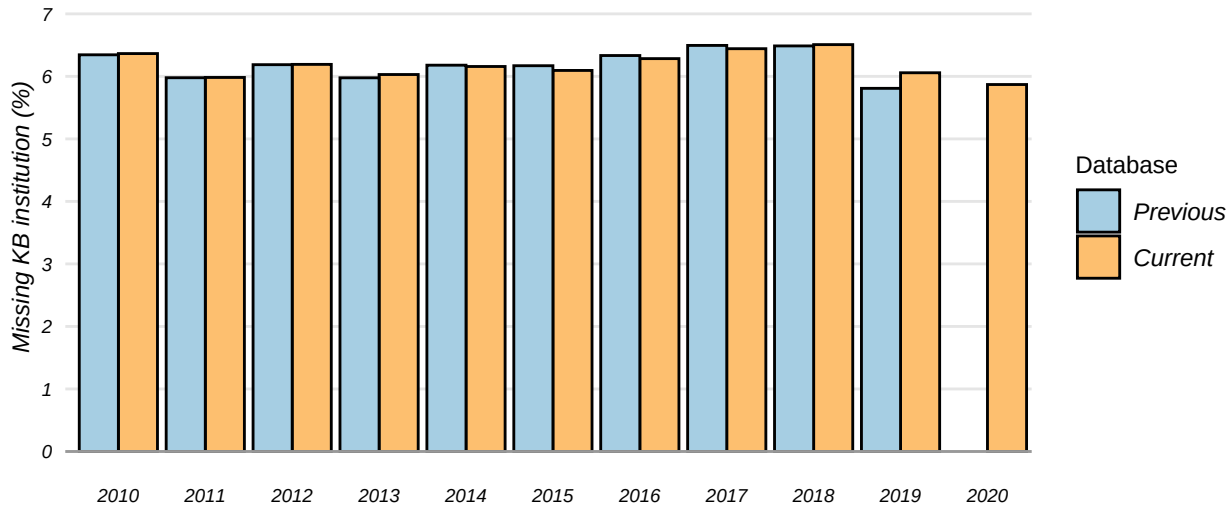

Figure 20: The number of German publications in each database that are missing a KB institution.

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 8 presents potentially new institutions – these had no publications in 2019 in the scopus_b_2020 database but more than five publications in 2020 in the scopus_b_2021 database. Conversely, Table 9 shows the institutions that had at least five publications in 2019 in the scopus_b_2020 database but no publications recorded in 2020 in the scopus_b_2021 database. We also highlight in Tables 10 and 11 the larger institutions (with at least 20 publications) that had a change in publication counts of more than 40% between 2019 and 2020 in the scopus_b_2020 and scopus_b_2021 databases.

Table 8: Institutions with more than 5 publications in 2020 in scopus_b_2021 that had no publications in 2019 in the scopus_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs
5555	Paracelsus Medizinische Privatuniversitat (PMU)	0	141
5352	Psychologische Hochschule Berlin	0	42
5569	Leipzig Heart Institute GmbH	0	42
5550	Institute for Advanced Sustainability Studies e.V. (IASS)	0	40
5529	HHL Leipzig Graduate School of Management	0	37
4736	IPPMed - Institut fur Pharmakologie und praventive Medizin GmbH	0	27
5543	Climate Analytics gGmbH	0	27
835	Bundesinstitut fur Bevolkerungsforschung	0	20
5544	Monasterium Laboratory - Skin & Hair research Solutions GmbH	0	17
5554	Fraunhofer-Einrichtung fur Wertstoffkreislaufe und Ressourcenstrategie IWKS	0	17
5546	WSG - Westdeutsche Studiengruppe GmbH	0	12
5547	Spectral Service AG	0	12
5572	GeneWerk GmbH	0	12
5468	CISPA ? Helmholtz-Zentrum fur Informationssicherheit	0	11
5545	SphingoTec GmbH	0	10
5551	Fraunhofer-Einrichtung fur Energieinfrastrukturen und Geothermie IEG	0	10
5535	Bard College Berlin ? A Liberal Arts University	0	9
5530	Max-Planck-Forschungsstelle fur die Wissenschaft der Pathogene	0	8
349	MTG Malteser Tragergesellschaft gemeinnutzige GmbH	0	7
1025	Max-Planck-Institut fur auslandisches und internationales Privatrecht	0	7
1504	EnBW Energie Baden-Wurttemberg AG	0	7
4176	MEDA Pharma GmbH	0	7
3769	Berliner Wasserbetriebe	0	6
5548	DERMATOLOGIKUM BERLIN Gemeinschaftspraxis GbR	0	6

Table 9: Institutions with no publications in 2020 in scopus_b_2021 that had more than 5 publications in 2019 in the scopus_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs

Table 10: Institutions with more than 20 publications in 2019 in the scopus_b_2020 that increased in publication counts by over 40% to 2020 in the scopus_b_2021 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
5252	Medizinische Hochschule Brandenburg Theodor Fontane	81	197	116	143.2
198	Senckenberg Forschungsinstitut und Naturmuseum Frankfurt	155	309	154	99.4
582	Hochschule Mittweida, University of Applied Sciences	22	42	20	90.9
602	Ernst-Abbe-Hochschule Jena ? University of Applied Sciences	28	51	23	82.1
4737	Institute for Advanced Sustainability Studies e.V.	55	98	43	78.2
1134	Fraunhofer-Institut fur Toxikologie und Experimentelle Medizin	47	83	36	76.6
5324	Mercator Research Institute on Global Commons and Climate Change	34	60	26	76.5
1116	Fraunhofer-Institut fur Umwelt-, Sicherheits- und Energietechnik	38	67	29	76.3
80	Private Universitat Witten/Herdecke gGmbH	460	806	346	75.2
179	ESCP Europe Wirtschaftshochschule Berlin e.V.	24	42	18	75.0
756	Forschungszentrum caesar	22	38	16	72.7
5348	Deutsches Zentrum fur Lungenforschung	163	280	117	71.8
1115	Fraunhofer-Institut fur Holzforschung	23	39	16	69.6
2145	Deutsches Forschungszentrum fur Kunstliche Intelligenz GmbH	51	86	35	68.6
609	Hochschule fur Technik und Wirtschaft Berlin	34	56	22	64.7
5478	Leibniz-Institut fur Werkstofforientierte Technologien - IWT	34	56	22	64.7
604	Hochschule fur Wirtschaft und Recht Berlin	41	66	25	61.0
566	Hochschule Pforzheim - Gestaltung, Technik, Wirtschaft und Recht	22	35	13	59.1

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
362	Klinikum St. Georg	29	46	17	58.6
1144	Fraunhofer-Institut fur Physikalische Messtechnik	22	34	12	54.5
4198	Hasso-Plattner-Institut fur Softwaresystemtechnik (HPI)	40	61	21	52.5
569	Hochschule Osnabruck	55	83	28	50.9
211	Senckenberg Museum fur Naturkunde Gorlitz	28	42	14	50.0
5368	Translational Lung Research Center Heidelberg	36	53	17	47.2
856	Bundesanstalt fur Arbeitsschutz und Arbeitsmedizin	28	41	13	46.4
452	St.-Antonius-Hospital Eschweiler	26	38	12	46.2
669	Hochschule Aalen - Technik und Wirtschaft	39	57	18	46.2
640	Hochschule fur nachhaltige Entwicklung Eberswalde	33	48	15	45.5
1032	Max-Planck-Institut fur Gesellschaftsforschung	22	32	10	45.5
675	WHU - Otto Beisheim School of Management	62	90	28	45.2
37	Leibniz-Institut fur Plasmaforschung und Technologie e.V.	90	130	40	44.4
652	Hochschule Bochum - University of Applied Sciences	25	36	11	44.0
5210	Berliner Institut fur Gesundheitsforschung	660	949	289	43.8
626	Hochschule Fulda - University of Applied Sciences	46	66	20	43.5
840	Bundesinstitut fur Risikobewertung (BfR)	174	248	74	42.5
177	Hertie School of Governance	50	71	21	42.0
5325	Auditory Valley	48	68	20	41.7
1153	Fraunhofer-Institut fur Keramische Technologien und Systeme	68	96	28	41.2
713	Institut fur Herzinfarktforschung (IHF)	32	45	13	40.6

Table 11: Institutions with more than 20 publications in 2019 in the scopus_b_2020 that decreased in publication counts by over 40% to 2020 in the scopus_b_2021 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
48	Leibniz-Institut fur Atmospharenphysik e.V. an der Universitat Rostock (IAP)	42	25	-17	-40.5
1476	GBG Forschungs GmbH	29	17	-12	-41.4
443	Berufsgenossenschaftliche Unfallklinik	31	18	-13	-41.9
495	Marienhospital Bottrop gGmbH	21	12	-9	-42.9
552	Technische Hochschule Wildau (FH)	42	24	-18	-42.9
1541	Daimler AG	48	27	-21	-43.8
1347	Miltenyi Biotec GmbH	31	17	-14	-45.2
646	Hochschule fur angewandte Wissenschaften Coburg	39	21	-18	-46.2
4165	Leibniz-Institut fur Wirtschaftsforschung Halle (IWH)	34	18	-16	-47.1
4164	Leibniz-Institut fur Deutsche Sprache (IDS)	25	13	-12	-48.0
726	FZI Forschungszentrum Informatik	24	12	-12	-50.0
1166	Fraunhofer-Institut fur Bauphysik	28	14	-14	-50.0
693	Laser Zentrum Hannover e.V. (LZH)	49	23	-26	-53.1
3765	Evangelisch-Freikirchliches Krankenhaus und Herzzentrum Brandenburg in Bernau	36	15	-21	-58.3
3836	SPECS Surface Nano Analysis GmbH	31	11	-20	-64.5
4163	ILS - Institut fur Landes- und Stadtentwicklungsforschung	21	7	-14	-66.7
4411	Klinikum Frankfurt Hochst GmbH	33	10	-23	-69.7
5203	Nanosystems Initiative Munich (NIM)	74	20	-54	-73.0

Authors: Median number of authors by Subject Area and discipline

The median 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 21 the median number of authors per discipline in 2019 in both databases, and in the right panel the median number of authors per discipline in 2019 in the scopus_b_2021 database compared to 2020 in the scopus_b_2021 database.

While little change is expected to be seen in the left-hand panel of Figure 21 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 median number of authors changed by more than 1, based on the right-hand panel of Figure 21, are shown in Table 12. Also, to assess trends over a longer time-series and the full range of authors, we present the percentage of publications in each quartile of the range of authors in each Subject Area over the most recent years of both databases in Figure 22.

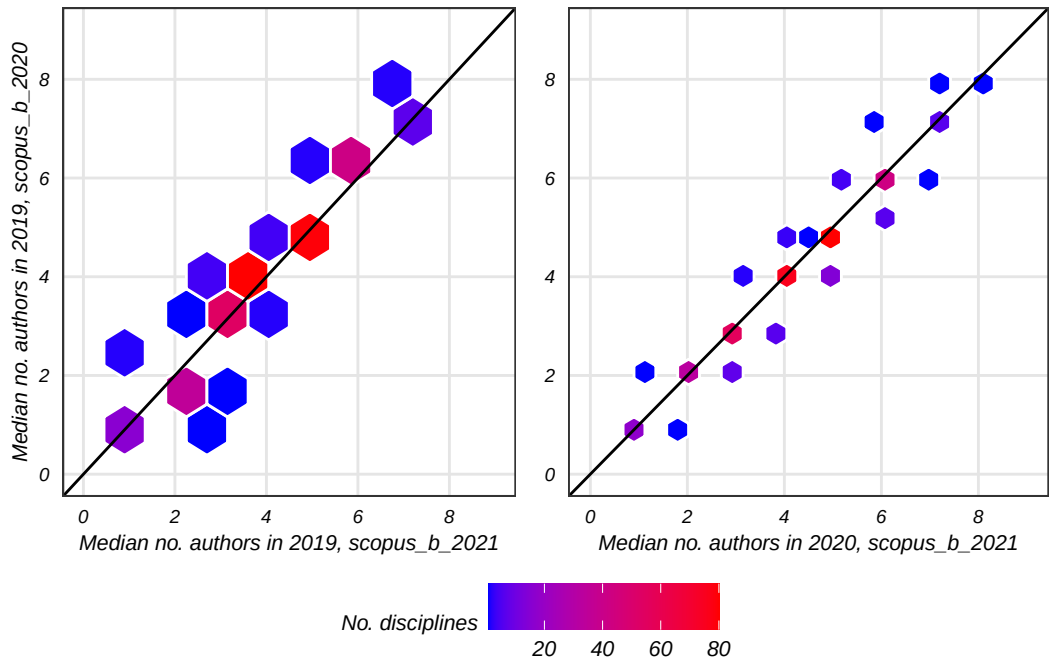

Figure 21: Median number of authors per discipline between databases, where colour denotes the number of disciplines with this combination of mean authors.

Table 12: Disciplines where the median number of authors changed by more than 1 between the last common year in the previous database and the latest year in the current database.

Discipline	Previous median authors	Current median authors	Diff.

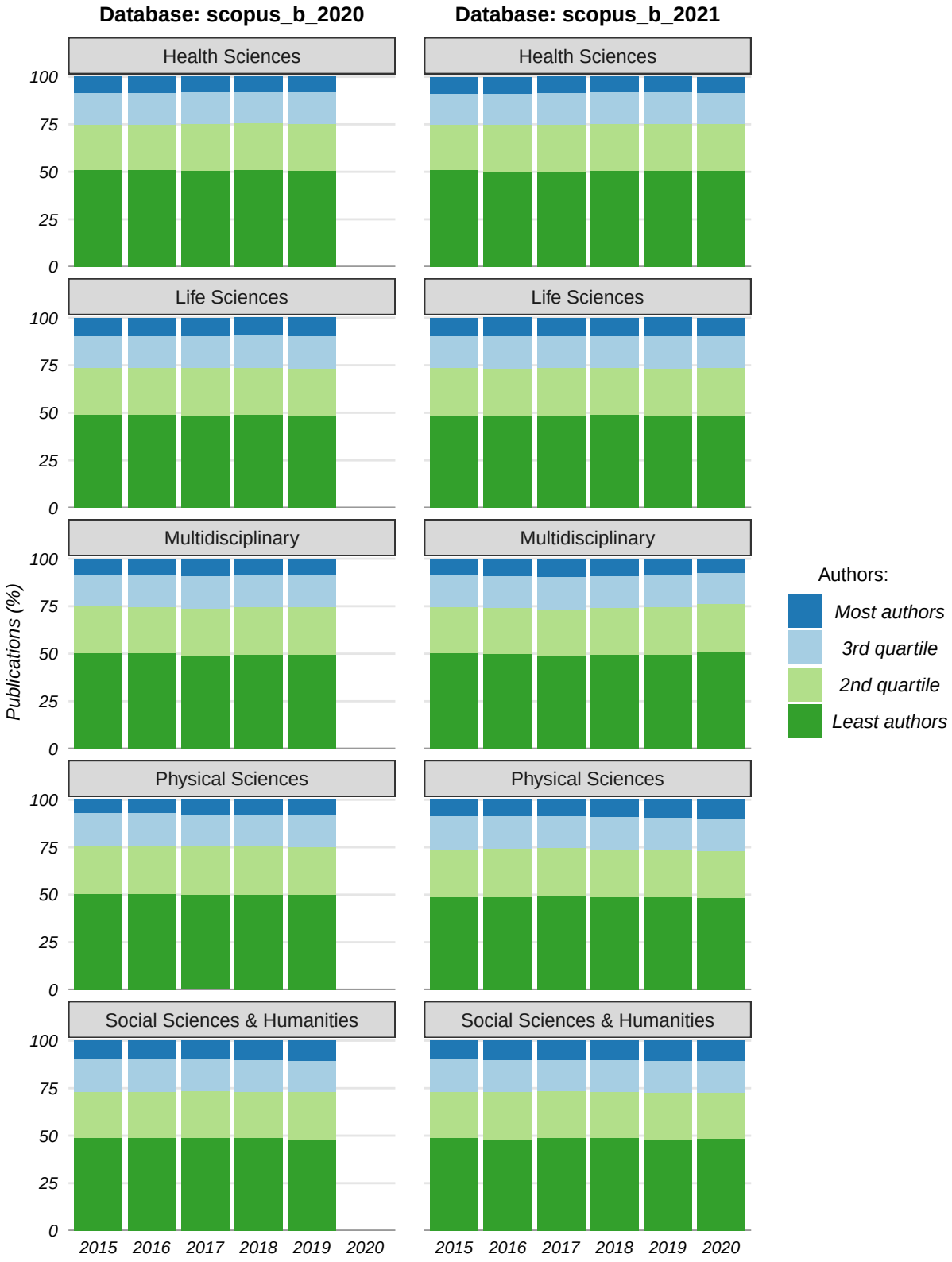

Figure 22: Distribution of publications over quartiles based on number of authors over time.

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 23 the percentage of references that are source items per discipline in 2019 in both databases, and in the right-hand panel the percentage of references that are source items per discipline in 2019 in the scopus_b_2021 database compared to 2020 in the scopus_b_2021 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 23, are shown in Table 13. Longer term trends can be seen in Figure 24 where we present the percentage of reference that are source items per Subject Area over the last ten common years of both databases.

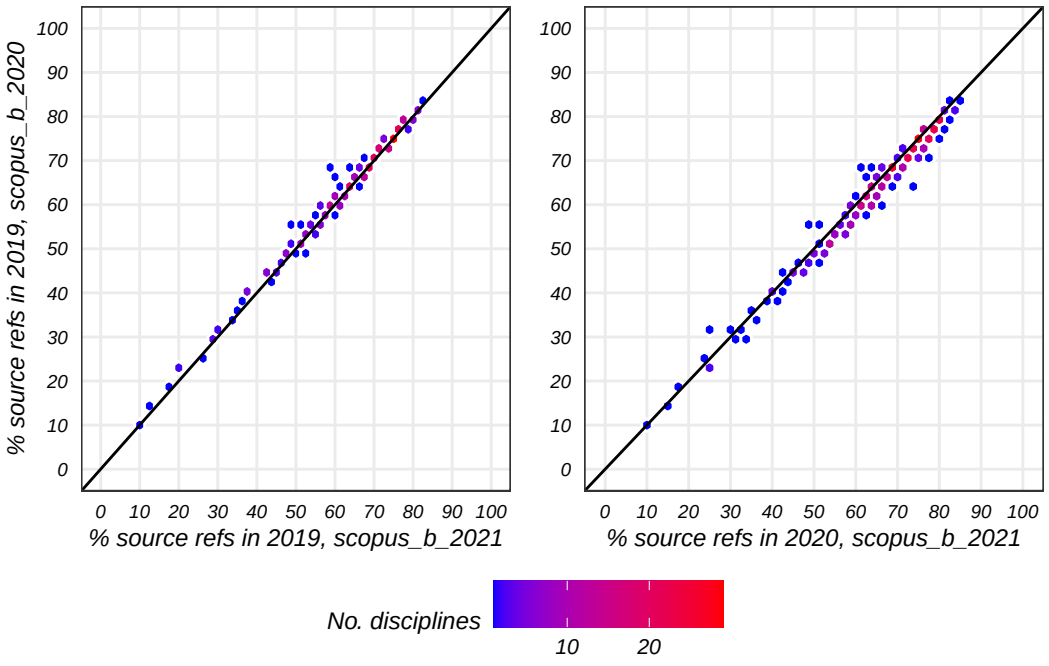

Figure 23: The percentage of cited items that are source items per discipline by database, where colour denotes the number of disciplines with this combination of source references.

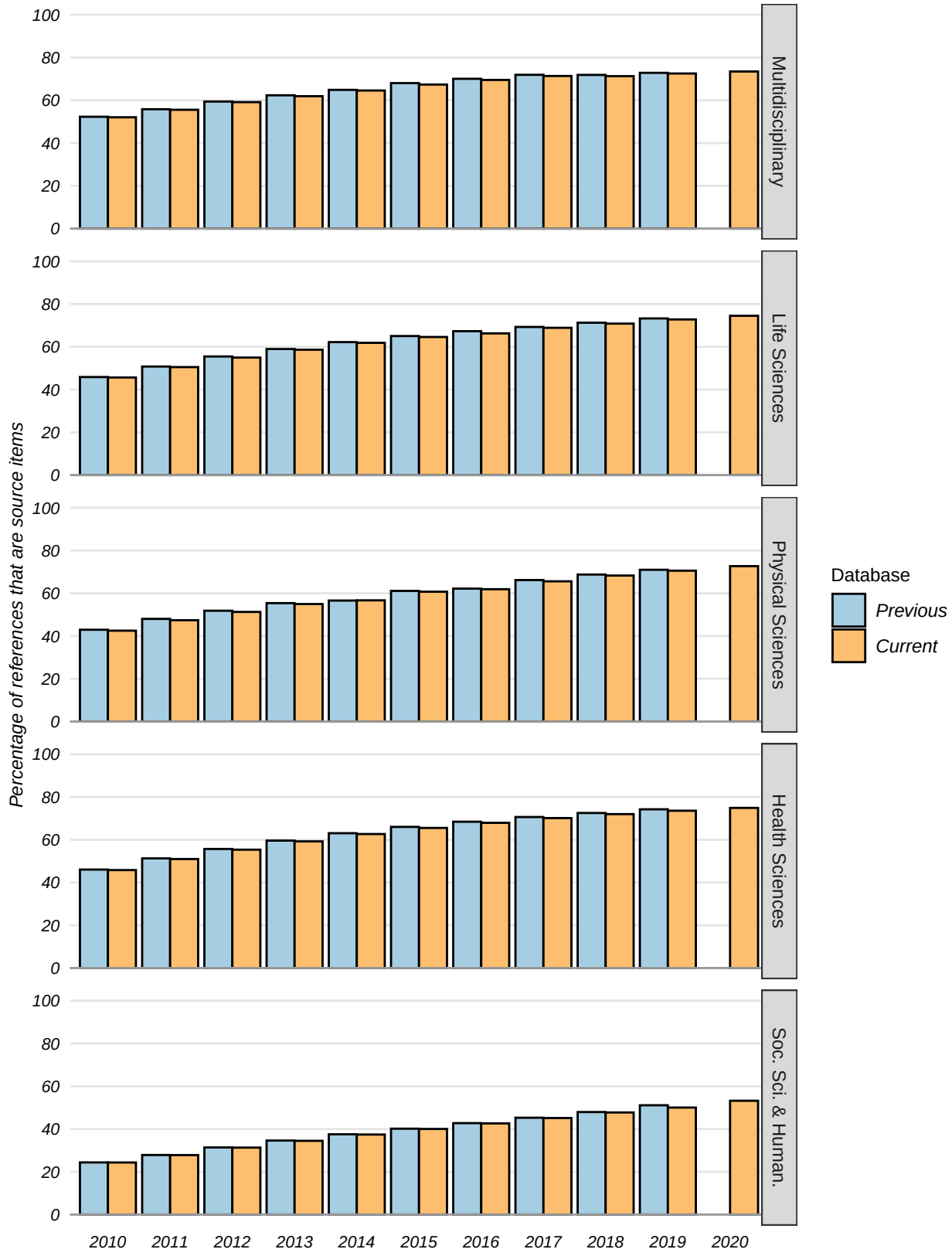

Figure 24: The percentage of references that are source items by Subject Area and database over time.

Table 13: Disciplines where the percentage of indexed references changed by 3 or more percentage points between 2019 in scopus_b_2020 and 2020 in scopus_b_2021.

Field	Previous no. refs.	Current no. refs.	Prvs % source	Crrnt % source	Change
Nurse Assisting	8,339	3,438	64.7	72.9	8.2
Physical & Theoretical Chemistry	22,375,036	24,040,257	69.9	76.4	6.5
Decision Sciences (misc.)	40,165	59,341	63.5	69.6	6.1
Pharmacology (nursing)	24,973	37,517	60.6	65.9	5.3
Gen. Arts & Humanities	463,379	443,901	29.8	34.8	5.0
Architecture	595,186	912,587	48.4	53.1	4.7
Pharmacology, Tox. & Pharmaceutics (misc.)	686,320	1,072,002	76.7	81.4	4.7
Mgmnt of Technology & Innovation	2,734,401	2,580,268	50.4	54.9	4.5
Business & International Mgmnt	3,134,149	3,445,371	50.6	54.7	4.1
Histology	1,841,023	2,670,544	72.7	76.8	4.1
Strategy & Mgmnt	4,935,497	6,070,314	58.9	62.9	4.0
Optometry	135,363	151,474	66.0	70.0	4.0
Gen. Physics & Astronomy	27,070,832	23,248,697	70.1	73.9	3.8
Review & Exam Preparation	16,640	14,370	63.1	66.9	3.8
Automotive Engineering	941,494	1,238,092	63.2	67.0	3.8
Human Factors & Ergonomics	512,407	598,547	53.8	57.5	3.7
Gen. Materials Science	37,948,743	45,391,140	75.8	79.4	3.6
Acoustics & Ultrasonics	1,230,762	1,351,450	67.4	71.0	3.6
Mgmnt Science & Operations Research	1,878,203	2,238,665	58.7	62.2	3.5
Industrial & Manufacturing Engineering	9,115,082	11,320,412	72.3	75.8	3.5
Materials Science (misc.)	1,525,587	2,091,742	68.0	71.5	3.5
Pharmacology	15,734,569	17,994,476	74.2	77.6	3.4
Environmental Engineering	9,755,621	11,519,511	70.2	73.5	3.3
Numerical Analysis	587,741	739,895	56.7	60.0	3.3
Waste Mgmnt & Disposal	8,211,821	10,234,266	71.8	74.9	3.1
Health Professions (misc.)	330,851	763,157	65.6	68.7	3.1
Gen. Mathematics	2,457,483	3,329,409	53.9	57.0	3.1
Mathematics (misc.)	210,215	254,054	45.1	48.2	3.1
Aerospace Engineering	2,303,340	2,555,511	61.0	64.1	3.1
Podiatry	37,104	49,352	62.5	65.6	3.1

Field	Previous no. refs.	Current no. refs.	Prvs % source	Crrnt % source	Change
Mechanics of Materials	14,591,077	16,695,316	72.6	75.6	3.0
Analysis	898,615	1,086,146	52.0	55.0	3.0
Physics & Astronomy (misc.)	24,622,941	12,879,831	68.3	64.9	-3.4
Paleontology	2,852,331	2,085,932	56.0	52.3	-3.7
Engineering (misc.)	11,606,787	5,541,915	66.1	61.3	-4.8
Mathematical Physics	1,634,876	2,040,738	54.8	49.6	-5.2
Medical Terminology	1,062	2,629	30.8	24.1	-6.7
Nuclear & High Energy Physics	29,667,313	11,539,191	68.3	61.3	-7.0

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.