

---

## **Quality assurance at the macro level: Comparing the current and previous Scopus snapshots**

---

Dimity Stephen, Stephan Stahlschmidt and Paul Donner

*December 2022*

**Editor:**

German Centre for Higher Education Research and Science Studies (DZHW) GmbH

Lange Laube 12 | 30159 Hannover | Germany | [info@dzhw.eu](mailto:info@dzhw.eu) | [www.dzhw.eu](http://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:**

Dr. habil. Thorsten Kowalke

**Registration Court:**

Amtsgericht Hannover | HRB 6489

VAT No.: DE291239300

December 2022

# Contents

<b>Motivation</b>	<b>1</b>
Set of indicators . . . . .	1
Set of entities . . . . .	2
Methodological details . . . . .	2
<b>Analysis</b>	<b>3</b>
Publication counts: Total, selected countries, German sectors, and Subject Areas . . . . .	3
Journals: Total indexed and the number added or removed . . . . .	6
Excellence Rates: Selected countries and German sectors . . . . .	7
Citations: Mean 3-year citations of articles and reviews by discipline . . . . .	9
Uncited articles and reviews: Percent by selected countries and German sectors . . . . .	15
Disciplines: Changes in discipline classification . . . . .	17
Disciplines: Changes in articles and reviews by discipline . . . . .	18
Disciplines: Percentage of publications not assigned to a discipline . . . . .	19
Metadata: Changes in pubyear, doctype, pubtype and items removed . . . . .	19
Metadata: Missing metadata variables . . . . .	21
Institution and country data: Number of articles and reviews with missing data . . . . .	22
German institutions: German publications missing from KB institution coding . . . . .	23
German institutions: Changes in whole counts of articles and reviews . . . . .	23
Authors: Median number of authors by Subject Area and discipline . . . . .	28
Source items: Percentage by Subject Area and discipline . . . . .	29

## 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 the KB's 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, and disciplines.

This report is not an exhaustive analysis of the databases' content, nor does it investigate any anomalies identified in the databases. However, this report probes the core variables fundamental to typical 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. Of particular relevance in this year's report is the transition from Oracle to PostgreSQL for the latest database. This report aims to provide users of the KB databases with an overview of any potential changes soon after the databases enter the production environment, so that these factors may be considered in analyses.

## Set of indicators

The indicators included in the report 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\_2021 and scp\_b\_202204 databases is 2020, 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 2018 for the scopus\_b\_2021 and scp\_b\_202204 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 compare the number of publications per discipline in 2018 from the scopus\_b\_2021 database against 2019 in the scp\_b\_202204 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 (Clinic), and combined Ressortforschung-Bund and Ressortforschung-Länder (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 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 primarily 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 [1] 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 (SA) classification. The ASJC is a fine-grained classification that allows changes in specific disciplines to be analysed. However, given it contains over 250 categories, it is sometimes useful to use higher level of aggregation to present an overview of the disciplines. As such, we also present some data on the SA classification. The SA consists of five broad groups: Health Sciences, Life Sciences, Physical Sciences, Social Sciences and Humanities, and Multidisciplinary.

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 SA.

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 scp\_b\_202204 database contains a greater number of publications for the most recent years than the scopus\_b\_2021 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. Further, documents can be assigned to multiple types in the current database, but only one in the previous database.

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.

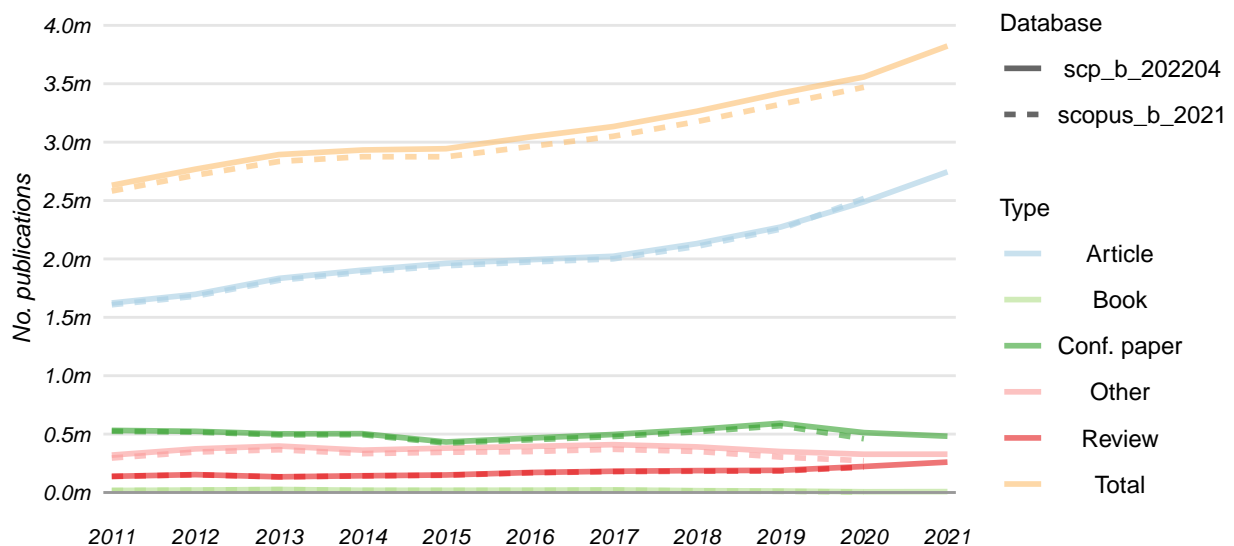


Figure 1: Number of documents in each database over time by type.

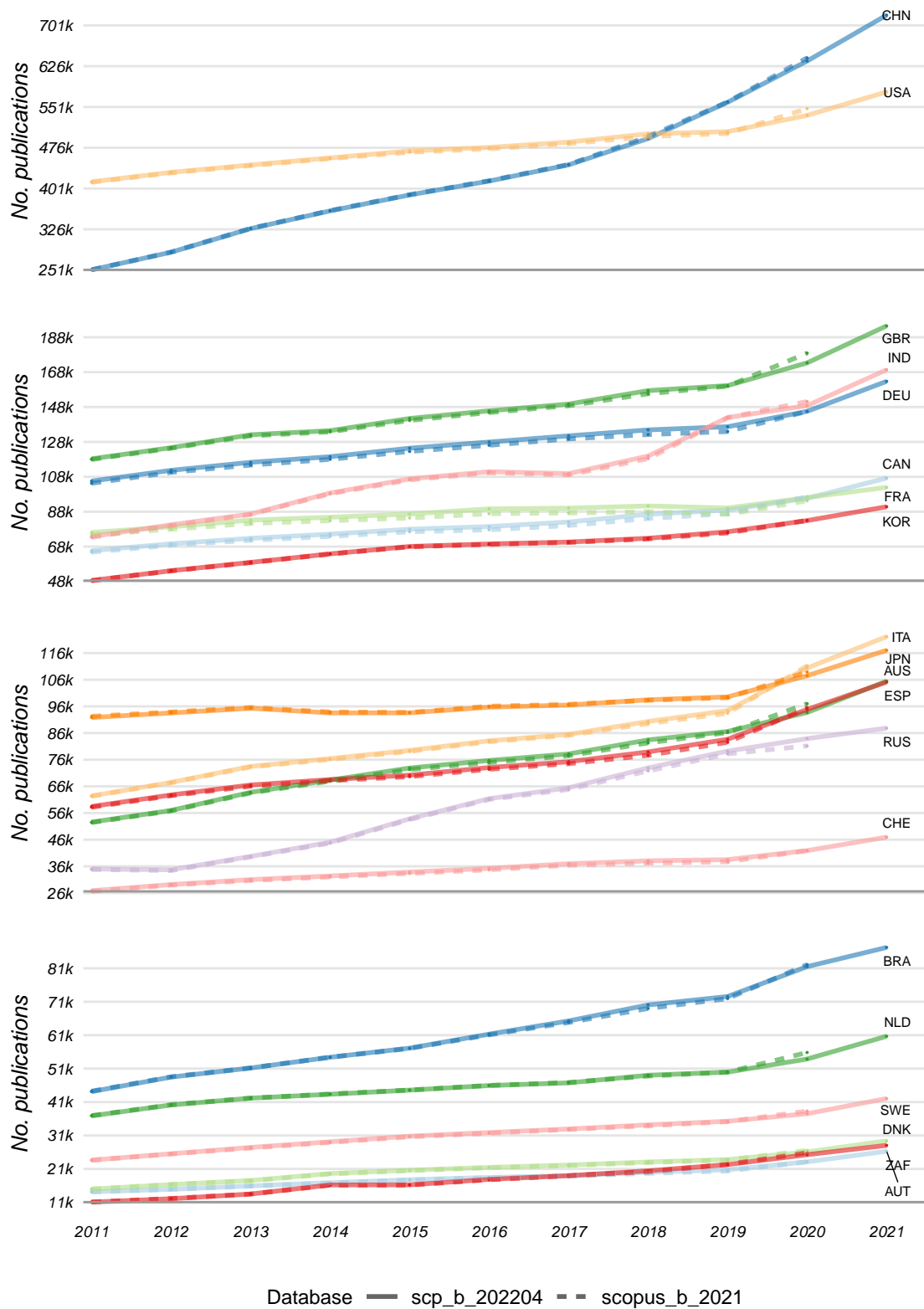


Figure 2: Whole counts of articles and reviews by country and database over time. Please note the panels' different axes.

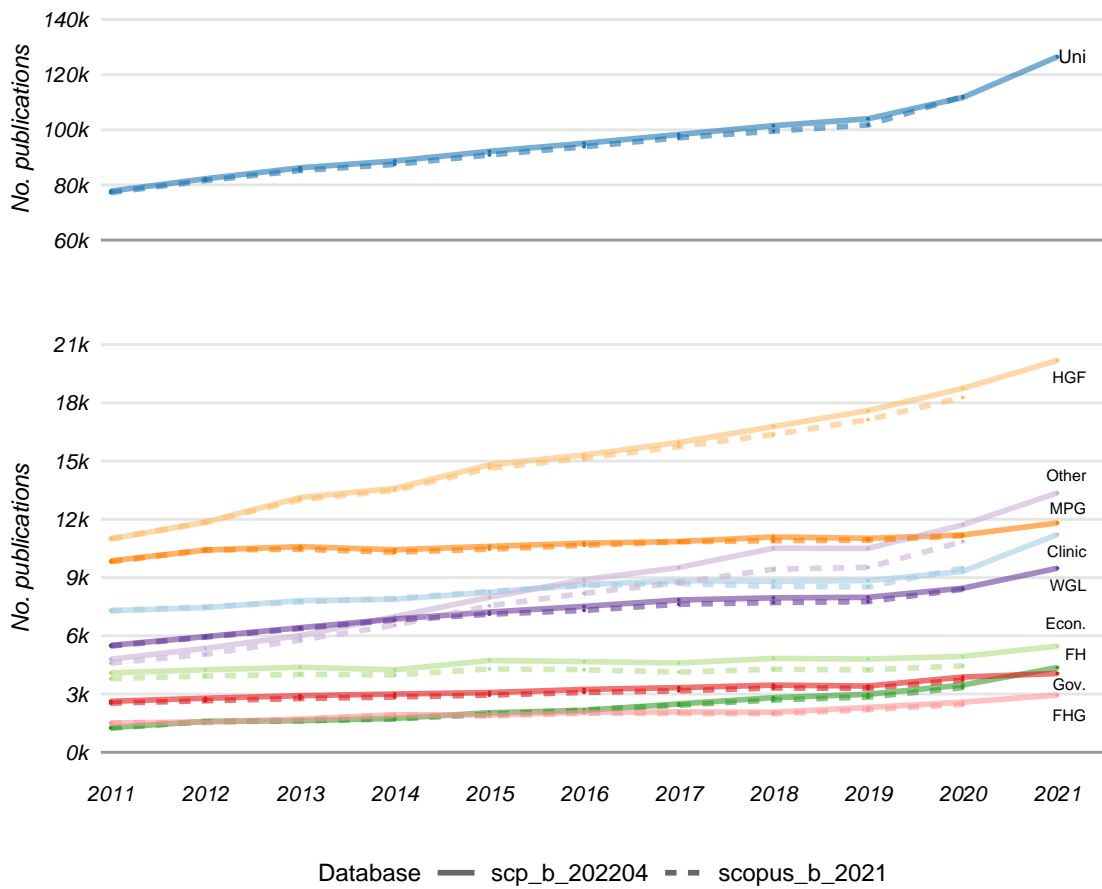


Figure 3: Whole counts of articles and reviews by German sector and database over time. Please note the panels' different axes.

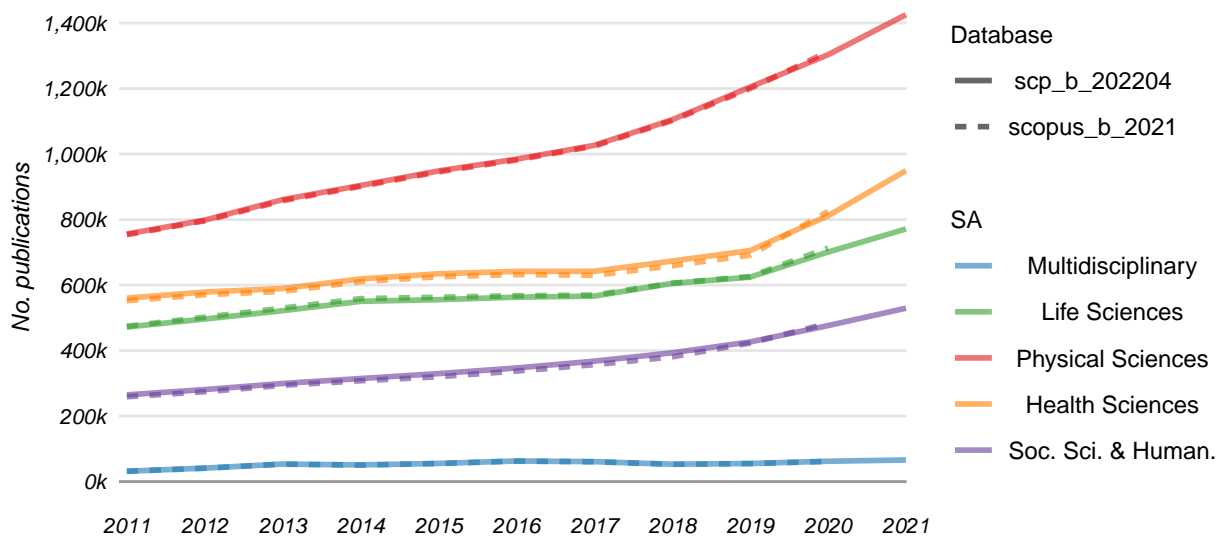


Figure 4: Whole counts of articles and reviews by SA and database over time.

## Journals: Total indexed and the number 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, particularly if changes are concentrated in specific disciplines. Figure 5 shows the total number of journals in each database over time, while Figure 6 shows the number of journals added and removed in each SA.

Changes in the journals indexed were identified by matching the titles of all journals indexed in 2020 in the `scopus_b_2021` database to those with 2021 content in the `scp_b_202204` database. Titles were used as all journals have titles recorded, while some journals are missing ISSNs. Titles in `scopus_b_2021` but not in `scp_b_202204` were considered removed, while titles in `scp_b_202204` but not in `scopus_b_2021` were considered added. In total, 2704 journals were added and 1285 were removed. These data may include a small number of journals that changed titles. Some double-counting of journals between SAs may also occur when a journal maps to two or more SAs.

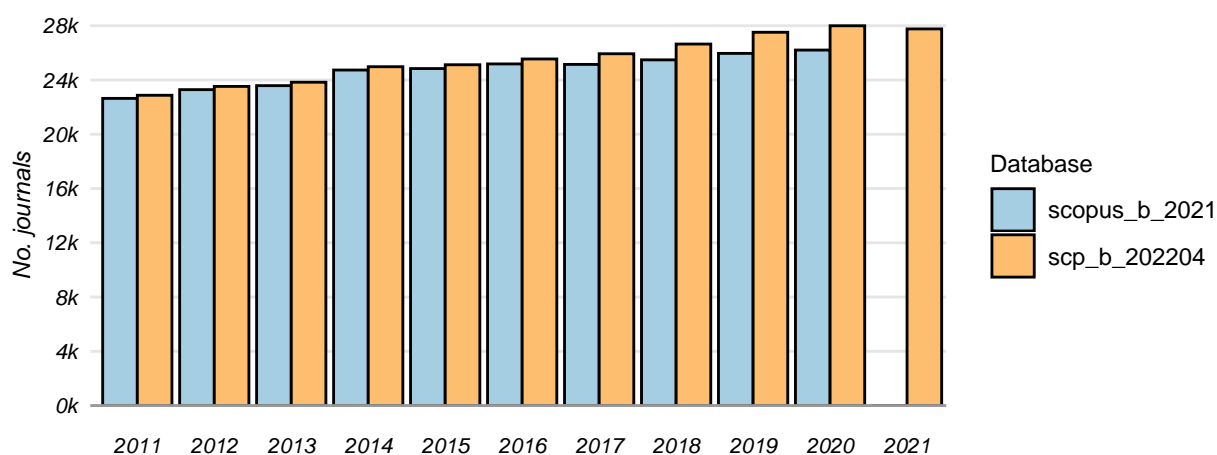


Figure 5: The number of journals indexed in each database over time.

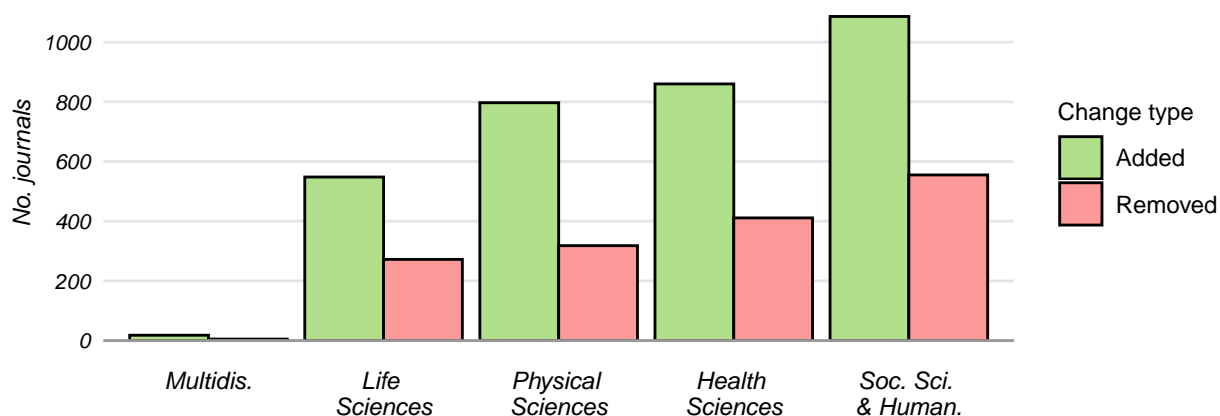


Figure 6: The number of journals added or removed between 2020 in `scopus_b_2021` and 2021 in `scp_b_202204` by SA.

### 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.

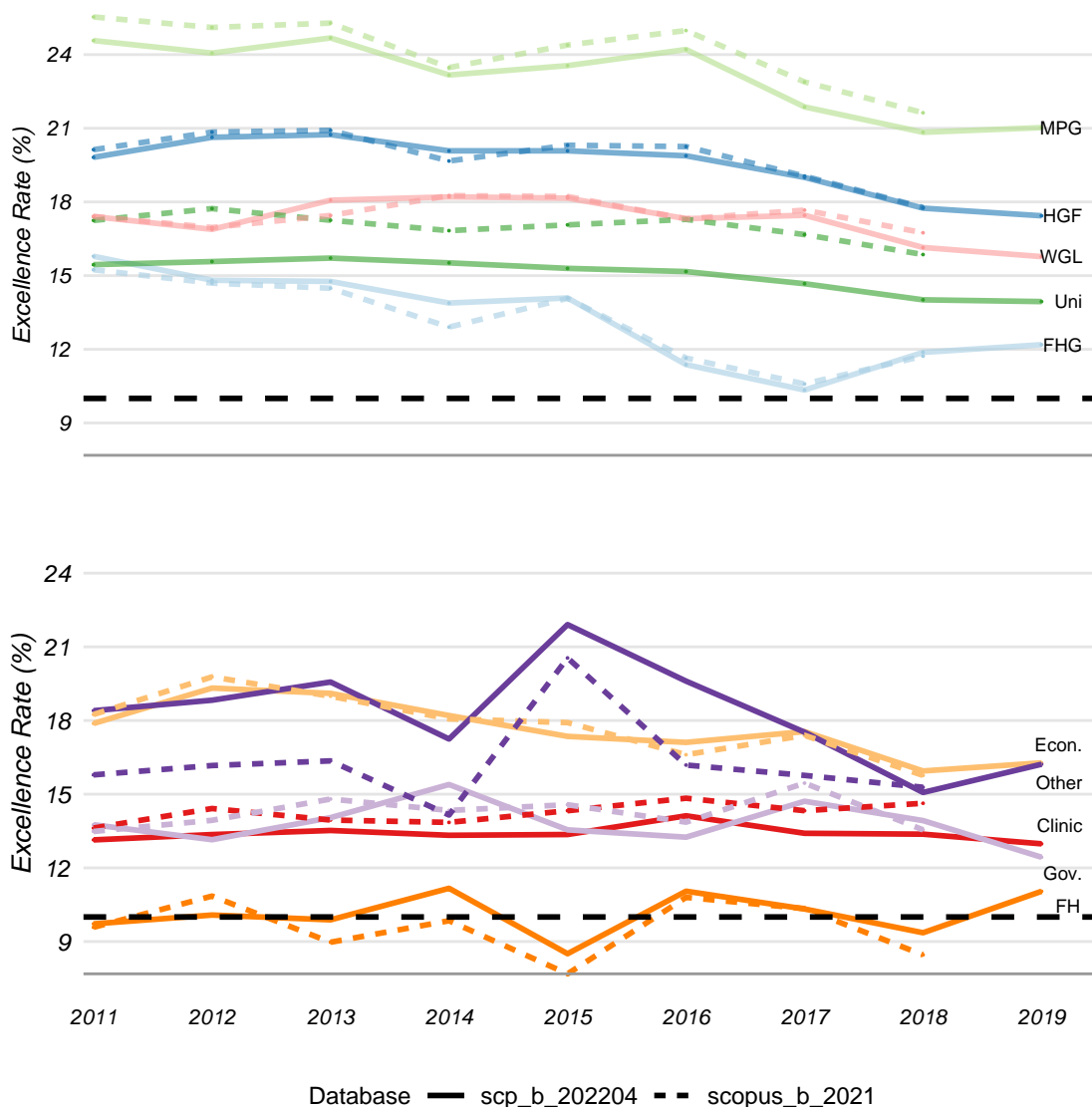


Figure 7: ERs by sector, based on whole counts. The black line is the expected 10% threshold.

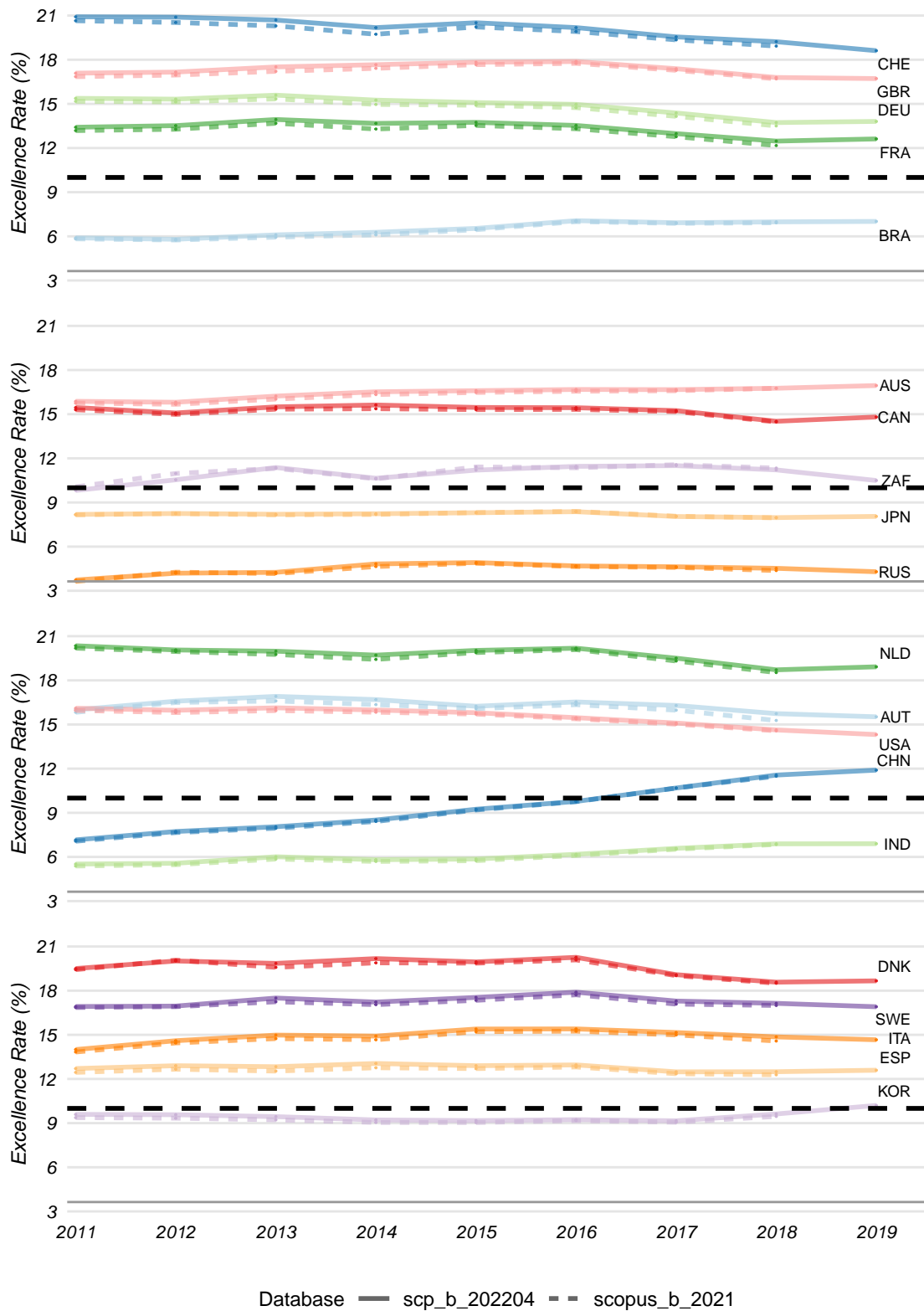


Figure 8: ERs for selected countries, based on whole counts. The black line is the expected 10% threshold.

## 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. We examine here in Figure 9 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 9 are shown in Tables 1 and 2, 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.

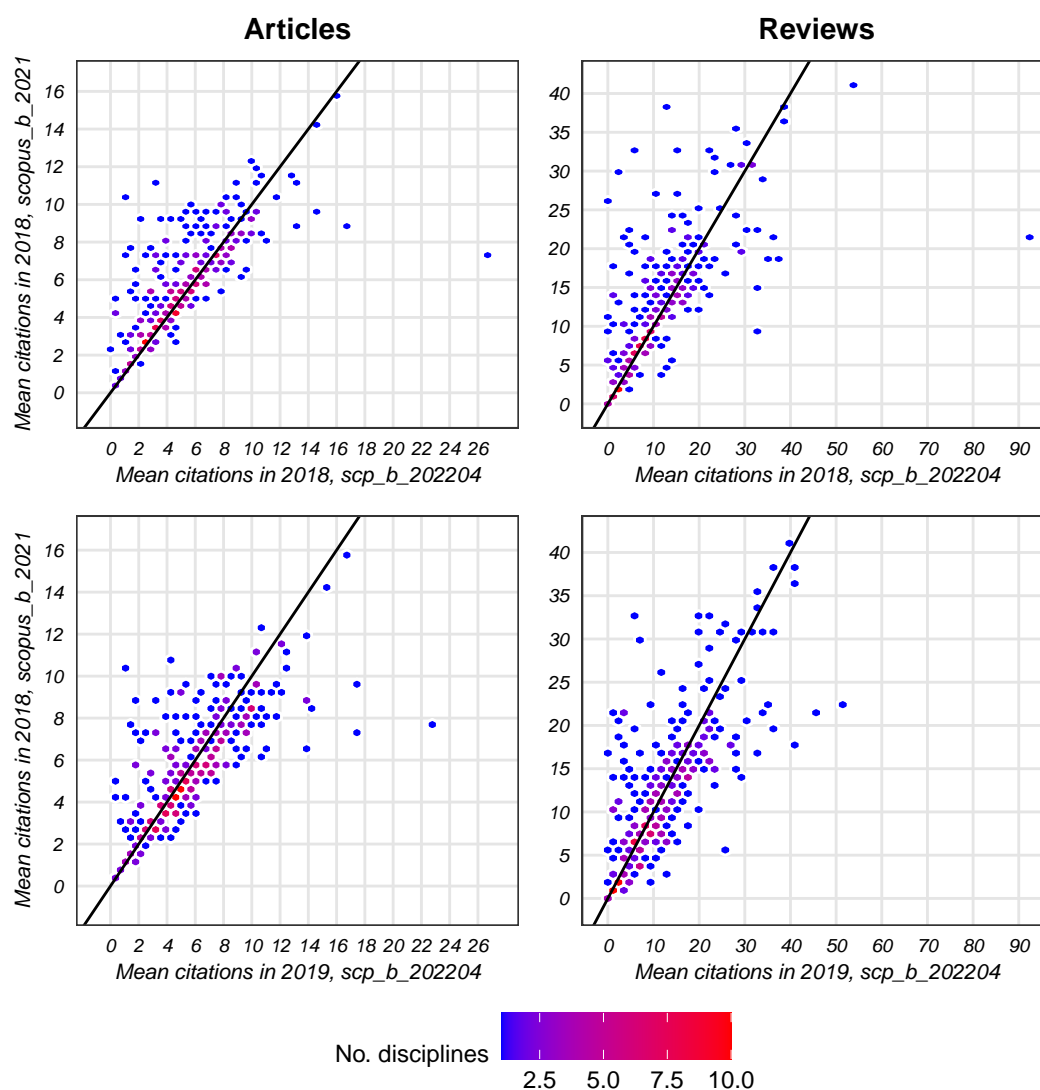


Figure 9: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

Table 1: Articles: Disciplines with a current MC3 of at least 1, where the MC3 decreased by over 20% or increased by over 50% between 2018 in scopus\_b\_2021 and 2019 in scp\_b\_202204, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. crnt pubs.	Perc. diff.
Biological Psychiatry	7.5	22.7	148	203.6
Energy (misc.)	7.1	17.4	129	145.6
Marketing	6.8	13.9	1895	105.3
Logic	3.2	6.1	545	89.2
Computer Science (misc.)	2.7	4.9	5328	83.5
Critical Care Nursing	2.3	4.2	350	81.0
Ecological Modeling	9.8	17.4	970	78.7
Soil Science	6.2	10.7	3410	73.7
Nursing (misc.)	3.0	5.0	114	67.4
Genetics	8.4	14.0	10317	66.3
Nutrition & Dietetics	6.5	10.8	638	65.7
Chemistry (misc.)	8.8	14.1	1285	61.2
Media Technology	5.7	9.2	577	59.4
Dentistry (misc.)	3.5	5.6	243	57.4
Building & Construction	9.0	13.8	4661	54.2
Periodontics	6.4	9.9	419	53.0
Oral Surgery	3.9	5.9	1816	50.3
Family Practice	2.3	3.4	1569	48.9
Computer Graphics & Computer-Aided Design	8.0	11.8	651	47.7
Business, Mgmt & Accounting (all)	4.1	6.0	5337	45.8
Numerical Analysis	5.1	7.4	2162	43.6
Neurology (clinical)	6.3	9.0	6571	42.7
Earth-Surface Processes	5.8	8.2	2271	40.6
Chemical Health & Safety	4.1	5.8	17	40.5
Tourism, Leisure & Hospitality Mgmt	7.9	11.1	1422	39.9
Materials Science (misc.)	4.0	5.5	2929	38.8
Histology	5.3	7.2	562	35.5
Complementary & Manual Therapy	2.7	3.7	574	35.0
Aging	7.8	10.4	2192	34.4
Business, Mgmt & Accounting (misc.)	4.1	5.6	1998	34.0
Genetics (clinical)	7.2	9.6	882	34.0
Developmental & Educational Psychology	5.3	7.1	3888	33.8
Archeology	2.7	3.6	165	33.3
Hardware & Architecture	5.9	7.9	1391	33.2
Earth & Planetary Sciences (all)	6.7	8.9	7757	32.7
Occupational Therapy	2.3	3.0	376	32.7
Life-span & Life-course Studies	4.9	6.4	196	31.9
Biotechnology	7.2	9.5	28908	31.7

Complementary & Alternative Medicine	3.4	4.5	2599	31.1
Arts & Humanities (all)	1.1	1.5	3818	30.9
Issues, Ethics & Legal Aspects	3.4	4.4	1349	30.2
Conservation	1.9	2.5	1776	30.2
Nurse Assisting	1.4	1.8	96	30.0
Chiropractics	2.6	1.8	212	-30.2
Spectroscopy	6.9	4.8	617	-31.3
Space & Planetary Science	9.3	6.2	712	-33.0
Mgmt, Monitoring, Policy & Law	7.9	5.2	161	-33.4
Safety Research	4.7	3.1	792	-34.6
Linguistics & Language	2.3	1.5	979	-35.4
Stratigraphy	7.2	4.6	48	-35.9
Speech & Hearing	3.7	2.3	303	-38.1
Applied Microbiology & Biotechnology	6.6	4.0	1111	-39.0
Earth & Planetary Sciences (misc.)	6.5	3.9	1137	-40.2
Computational Theory & Mathematics	7.9	4.6	252	-41.9
Health Professions (all)	4.0	2.3	91	-42.2
Geometry & Topology	3.3	1.9	1252	-43.3
Law	2.9	1.6	5086	-43.7
Agricultural & Biological Sciences (all)	5.4	3.0	7095	-44.9
Human-Computer Interaction	9.4	5.1	720	-45.2
Electrochemistry	9.3	4.9	1329	-47.2
Computer Vision & Pattern Recognition	9.5	5.0	410	-47.4
Mechanical Engineering	7.9	4.2	9833	-47.4
Engineering (all)	5.3	2.7	24155	-49.9
Industrial & Manufacturing Engineering	9.2	4.4	1925	-51.9
Endocrine & Autonomic Systems	7.4	3.5	68	-52.2
Nuclear Energy & Engineering	7.3	3.5	2860	-52.3
Library & Information Sciences	5.0	2.3	1935	-55.3
LPN & LVN	2.6	1.1	56	-56.9
Biochemistry, Genetics & Molecular Biology (misc.)	10.9	4.2	168	-61.6
History & Philosophy of Science	3.2	1.2	384	-61.7
Information Systems & Mgmt	8.8	3.1	81	-64.3
Metals & Alloys	7.4	2.6	2101	-65.2
Advanced & Specialized Nursing	6.7	2.2	667	-67.4
Pharmacology, Toxicology & Pharmaceutics (misc.)	5.7	1.6	1044	-71.0
Economic Geology	5.6	1.6	40	-71.0
Physiology (medical)	7.1	1.8	127	-75.0
Surfaces, Coatings & Films	9.0	1.7	693	-80.8
Materials Chemistry	7.8	1.4	125	-81.9

Table 2: Reviews: Disciplines with a current MC3 of at least 3, where the MC3 decreased by over 20% or increased by over 60% between 2018 in scopus\_b\_2021 and 2019 in scp\_b\_202204, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. crnt pubs.	Perc. diff.
Economics, Econometrics & Finance (misc.)	2.2	9.8	31	346.7
Mgmt Information Systems	6.2	25.7	86	315.7
Optometry	3.2	12.5	2	286.0
Archeology	1.3	4.3	32	218.3
Industrial Relations	1.6	4.1	62	164.9
Nursing (misc.)	3.9	9.5	22	142.9
Mathematics (misc.)	1.8	4.2	36	131.4
Electrochemistry	22.6	51.2	30	126.4
Accounting	5.0	11.4	137	126.3
Biological Psychiatry	18.3	41.0	22	124.1
Arts & Humanities (misc.)	3.3	7.5	183	123.8
Economics, Econometrics & Finance (all)	1.6	3.5	130	122.2
Marketing	6.4	14.1	32	120.7
Soil Science	14.3	30.3	59	112.1
Building & Construction	21.6	45.1	155	108.9
Media Tech.	3.6	7.6	16	108.8
Applied Psychology	9.7	19.9	212	105.7
Hardware & Architecture	16.7	33.9	48	102.3
Fundamentals & Skills	4.8	9.4	20	97.0
Tourism, Leisure & Hospitality Mgmt	7.4	14.5	27	96.9
Developmental & Educational Psychology	6.8	13.1	178	93.6
Automotive Engineering	14.8	28.6	71	93.6
Embryology	8.6	16.4	18	91.8
Statistical & Nonlinear Physics	7.0	13.1	85	87.2
Molecular Biology	19.4	35.7	1346	84.5
Business & International Mgmt	4.9	9.0	610	82.0
Plant Science	13.2	23.4	647	76.8
Issues, Ethics & Legal Aspects	2.7	4.8	192	75.8
Business, Mgmt & Accounting (misc.)	3.9	6.8	94	75.5
Small Animals	3.5	6.1	185	75.5
Agricultural & Biological Sciences (misc.)	6.9	12.0	77	74.0
Research & Theory	1.8	3.1	16	72.7
Complementary & Manual Therapy	4.1	6.7	113	64.7
Life-span & Life-course Studies	6.1	10.0	2	64.1
Physical & Theoretical Chemistry	21.0	34.2	512	63.3
Chemical Engineering (misc.)	17.0	27.7	137	62.9
Nutrition & Dietetics	14.1	11.2	64	-20.4
Cellular & Molecular Neuroscience	21.2	16.8	272	-20.8

Emergency Nursing	5.0	4.0	65	-20.8
Family Practice	4.8	3.8	139	-22.3
Condensed Matter Physics	22.5	17.4	288	-22.6
Engineering (all)	15.9	12.3	412	-22.7
Environmental Science (misc.)	11.4	8.6	228	-24.7
Transplantation	8.3	6.2	169	-24.7
Mathematics (all)	10.8	8.1	246	-24.7
Computer Science Applications	18.2	13.6	240	-25.5
Insect Science	10.9	8.1	153	-25.7
Agricultural & Biological Sciences (all)	11.5	8.5	415	-25.8
Electrical & Electronic Engineering	20.1	14.8	332	-26.4
Immunology & Microbiology (misc.)	7.1	5.2	54	-26.6
Computer Science (misc.)	11.8	8.7	105	-26.8
Immunology & Microbiology (all)	16.4	12.0	164	-27.0
Inorganic Chemistry	20.4	14.8	151	-27.4
Geochemistry & Petrology	13.6	9.9	45	-27.6
Energy Engineering & Power Tech.	27.3	19.6	229	-28.1
Nuclear & High Energy Physics	12.8	9.2	519	-28.6
Mgmt of Tech. & Innovation	7.8	5.5	11	-29.2
Histology	13.0	9.2	100	-29.2
Atomic & Molecular Physics, & Optics	23.9	16.1	364	-32.4
Developmental Neuroscience	14.8	9.7	258	-34.4
Pollution	32.5	21.1	244	-35.2
Drug Discovery	13.8	8.8	666	-36.3
Safety Research	10.5	6.6	63	-37.1
Signal Processing	30.9	18.9	125	-38.7
Health Professions (all)	8.5	5.1	7	-39.5
Mechanics of Materials	32.8	19.5	230	-40.7
Instrumentation	17.5	10.3	204	-41.3
Earth & Planetary Sciences (misc.)	7.0	3.9	9	-44.2
Advanced & Specialized Nursing	9.5	5.0	80	-47.6
Computers in Earth Sciences	12.1	6.1	8	-49.2
Speech & Hearing	6.6	3.4	14	-49.2
Metals & Alloys	14.0	6.2	33	-55.9
Applied Mathematics	13.2	5.5	14	-58.2
Mechanical Engineering	26.5	10.9	300	-59.0
Biochemistry, Genetics & Molecular Biology (misc.)	22.3	9.0	48	-59.8
Mgmt Science & Operations Research	16.5	6.0	10	-63.7
Spectroscopy	19.5	6.3	15	-67.9
Health Professions (misc.)	15.4	4.0	20	-73.7
Physiology (medical)	14.2	3.6	41	-75.0
Surfaces, Coatings & Films	29.9	7.3	12	-75.4
Control & Optimization	17.3	4.0	3	-76.9

---

Acoustics & Ultrasonics	21.3	4.1	19	-80.7
Nuclear Energy & Engineering	32.7	5.1	101	-84.3
Engineering (misc.)	21.1	3.2	29	-84.7

---

### 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 the database provider 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 10 and 11 the percentage of articles and reviews per German sector and selected country that remained uncited 3 years after they were published.

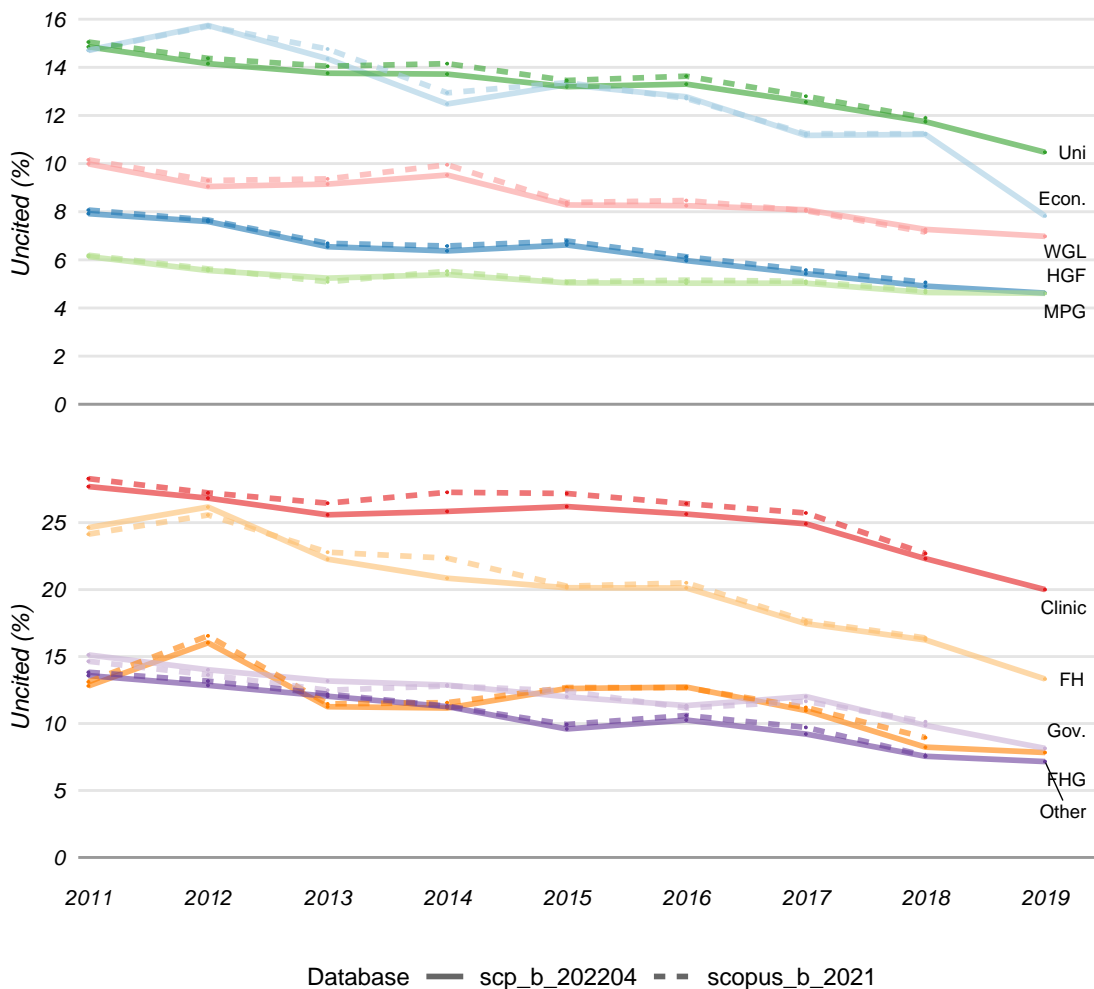


Figure 10: The percentage of uncited publications in each database over time by German sector.

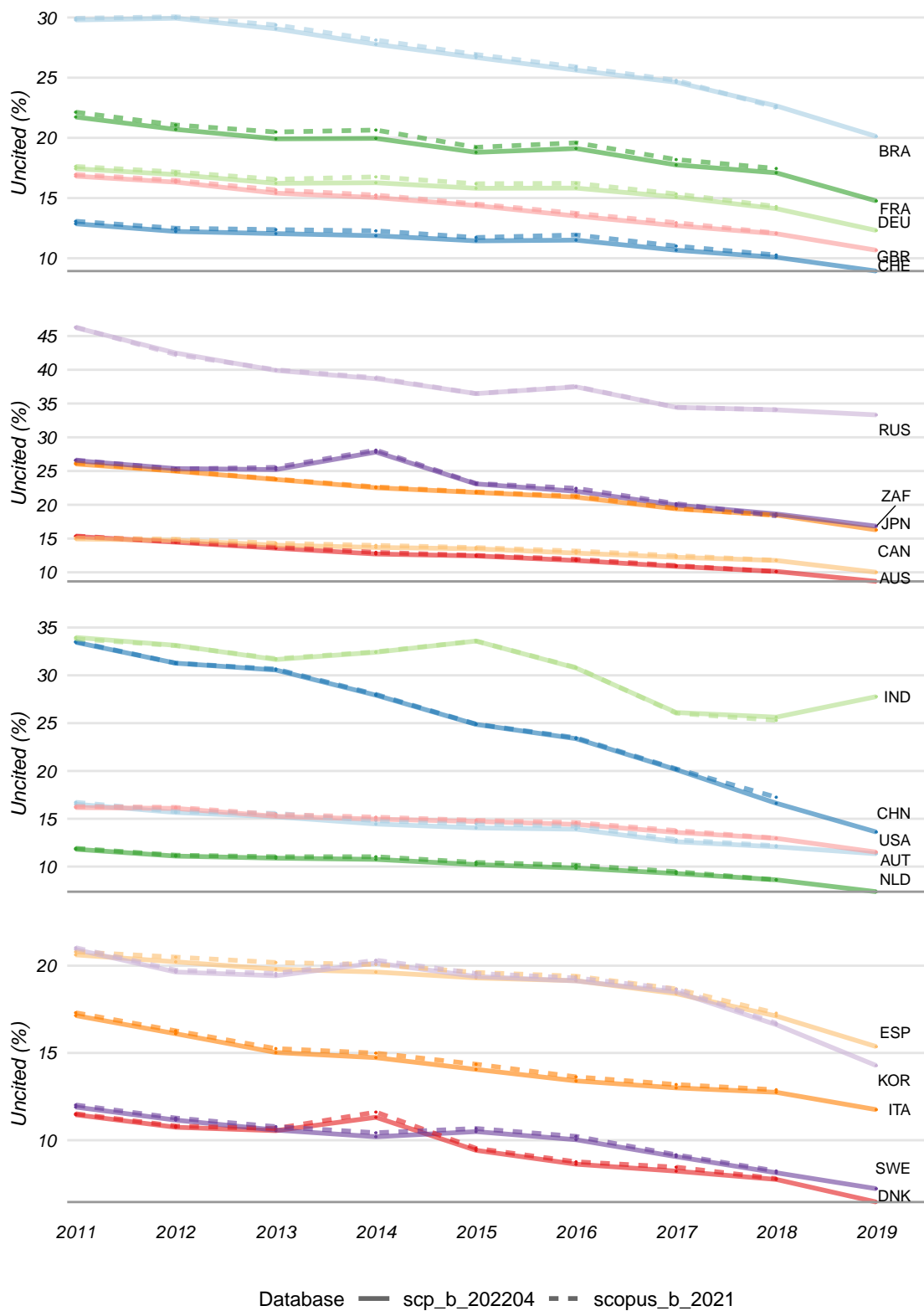


Figure 11: The percentage of uncited publications in each database over time by selected countries.

## Disciplines: Changes in discipline classification

This section shows in Table 3 any changes that have been made to Scopus' discipline classification, the 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 12 shows the number of publications assigned to specific disciplines identified to have changed in recent versions of the database.

Table 3: Changes in the ASJC discipline classification structure between the previous and current databases

Code	Classification	Previous pubs	Current pubs
NA		NA	4
3615	Respiratory Care	NA	39
2414	NA	54	NA
3330	NA	38	NA
3606	Medical Assisting and Transcription	46	NA

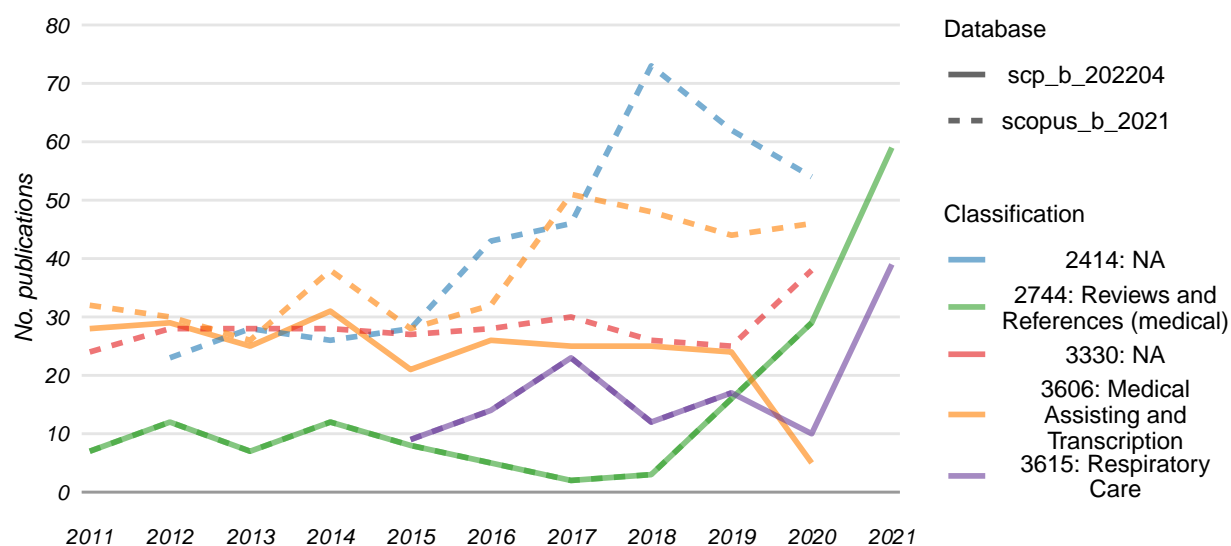


Figure 12: Time-series of disciplines known to have changed in recent versions of the databases. Dashed lines show the previous database and full lines show the current database.

## 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 may 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 13 the 40 disciplines with the highest percentage increases and decreases in publication counts between 2020 in scopus\_b\_2021 and 2021 in scp\_b\_202204. 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.

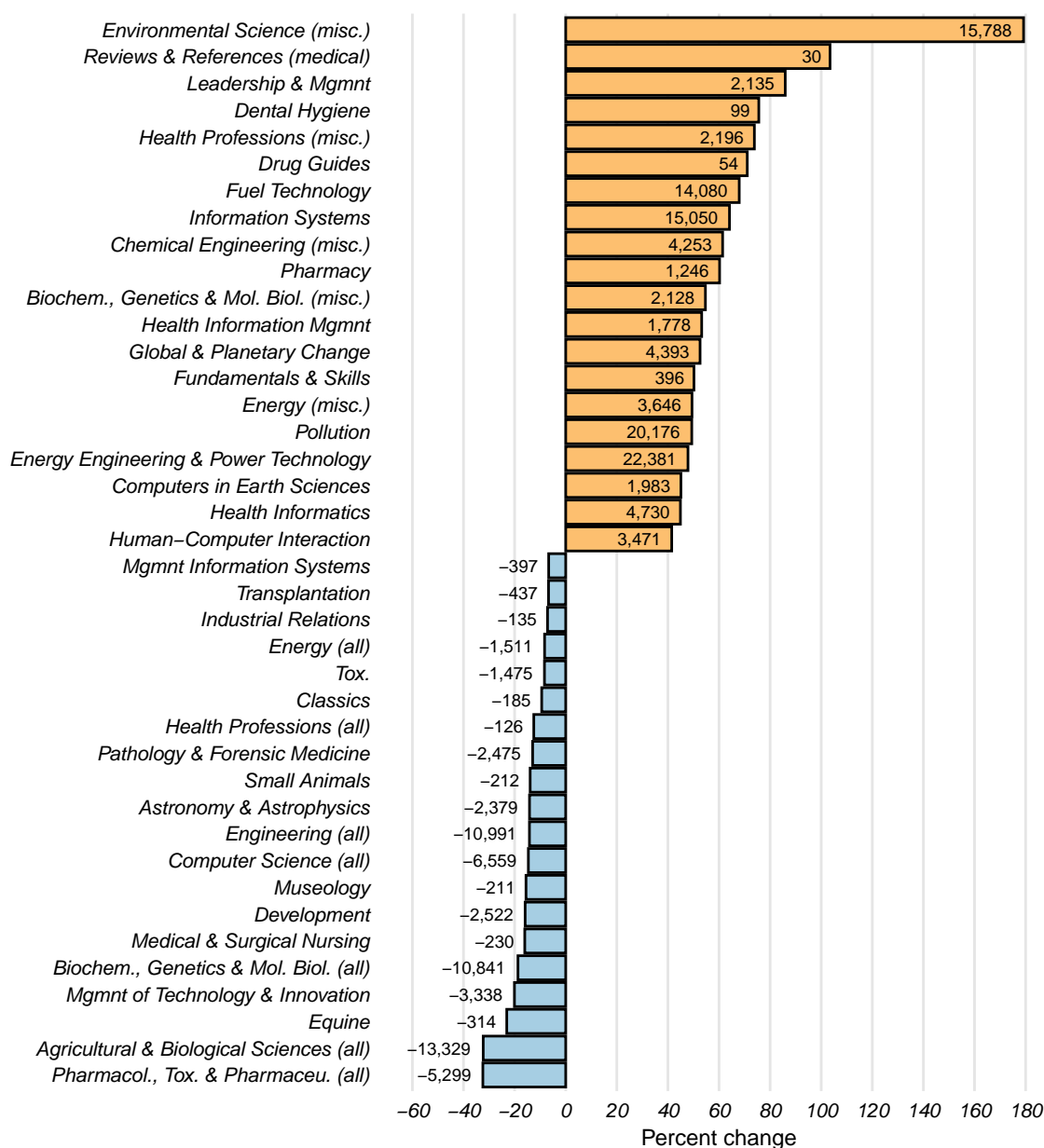


Figure 13: The 40 disciplines with the highest percentage change in publication counts between 2020 in scopus\_b\_2021 and 2021 in scp\_b\_202204, with numerical difference in counts.

## Disciplines: Percentage of publications not assigned to a discipline

Figure 14 shows the percentage of publications in each database that were not assigned to a discipline over the previous 11 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.

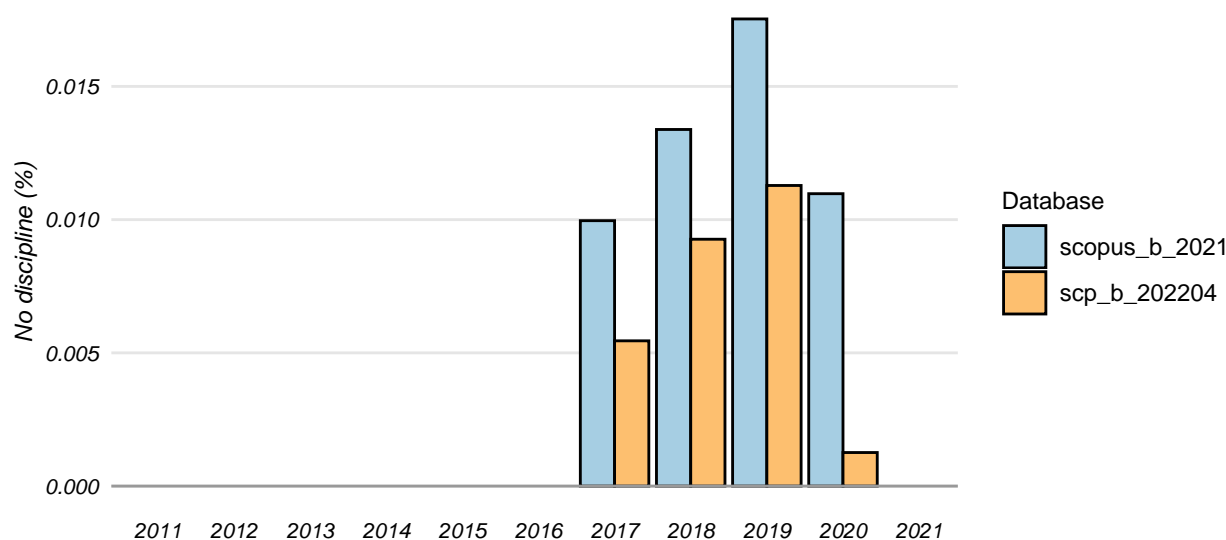


Figure 14: The percentage of publications in each database that do not have a discipline classification.

## 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\_2021 database but not in the scp\_b\_202204 database (removed items), and items present in the scp\_b\_202204 database but not in the scopus\_b\_2021 database (added items). 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 the database provider updates or corrects items. However, changes to or removal of a large number of items may require investigation. Notably, differences in document type may stem from the assignment of documents to multiple types in PostgreSQL compared to just one type in Oracle. Also, the documents examined are not restricted to articles and reviews, but any document type.

We identified changes in the metadata of in-scope items by first matching items between the scopus\_b\_2021 and scp\_b\_202204 databases using the UT\_EID identifier and then calculating the number of items that were added, removed, or had different metadata. The results are shown in Table 4.

Table 4: The number of items with changes in metadata between scopus\_b\_2021 and scp\_b\_202204.

Crrnt year	Prvs year	Diff. year	Diff. pubtype	Diff. doctype	Added	Removed
2016	2016	NA	102621	3567	NA	NA
2016	2017	91	1	15	NA	NA
2016	2018	6	NA	4	NA	NA
2016	2019	27	NA	NA	NA	NA
2016	2020	18	NA	NA	NA	NA
2016	NA	NA	NA	NA	47714	NA
2017	2016	42	NA	7	NA	NA
2017	2017	NA	104053	4084	NA	NA
2017	2018	167	33	11	NA	NA
2017	2019	29	1	NA	NA	NA
2017	2020	35	NA	NA	NA	NA
2017	NA	NA	NA	NA	45024	NA
2018	2016	19	NA	3	NA	NA
2018	2017	384	1	193	NA	NA
2018	2018	NA	118391	6327	NA	NA
2018	2019	130	1	21	NA	NA
2018	2020	669	1	2	NA	NA
2018	NA	NA	NA	NA	64625	NA
2019	2016	9	NA	NA	NA	NA
2019	2017	32	NA	16	NA	NA
2019	2018	883	2	413	NA	NA
2019	2019	NA	115730	3775	NA	NA
2019	2020	946	31	36	NA	NA
2019	NA	NA	NA	NA	99105	NA
2020	2016	6	NA	3	NA	NA
2020	2017	4	1	3	NA	NA
2020	2018	236	5	125	NA	NA
2020	2019	5546	2	107	NA	NA
2020	2020	NA	118750	5084	NA	NA
2020	NA	NA	NA	NA	202924	NA
NA	2016	NA	NA	NA	NA	7423
NA	2017	NA	NA	NA	NA	8934
NA	2018	NA	NA	NA	NA	15820
NA	2019	NA	NA	NA	NA	40672
NA	2020	NA	NA	NA	NA	155970

### Metadata: Missing metadata variables

Figure 15 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.

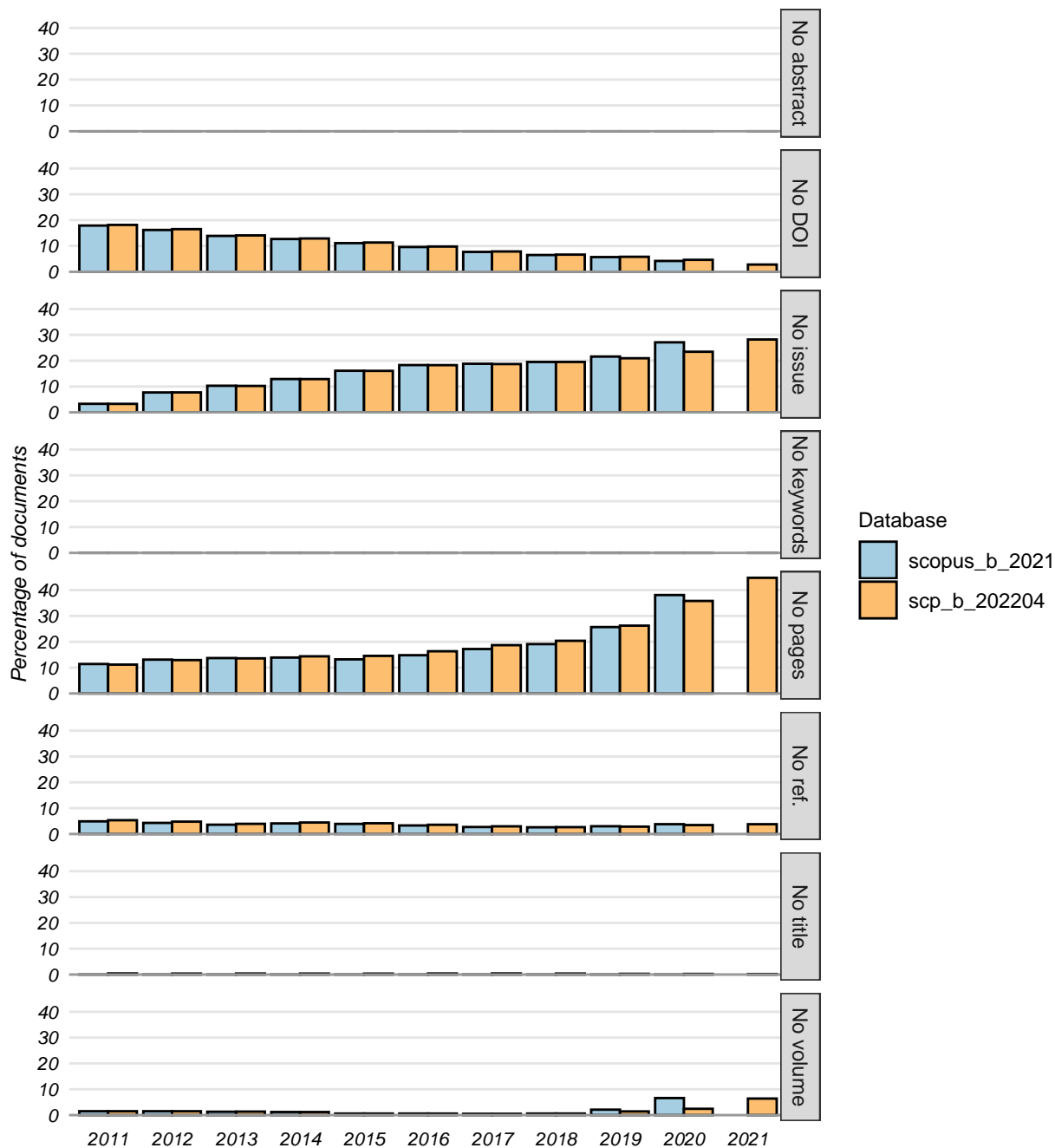


Figure 15: The percentage of items with missing metadata over time by database.

### 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 in `scopus_b_2021` and in the `items_affiliations` table in `scp_b_202204`. We have operationalised missing institution information here as publications that appear in the Items table but have no corresponding information in these affiliation tables. We present in the top panel of Figure 16 the number of items in each database between 2011 and 2021 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 16. Large disparities between the databases or substantial increases in missing information should be investigated.

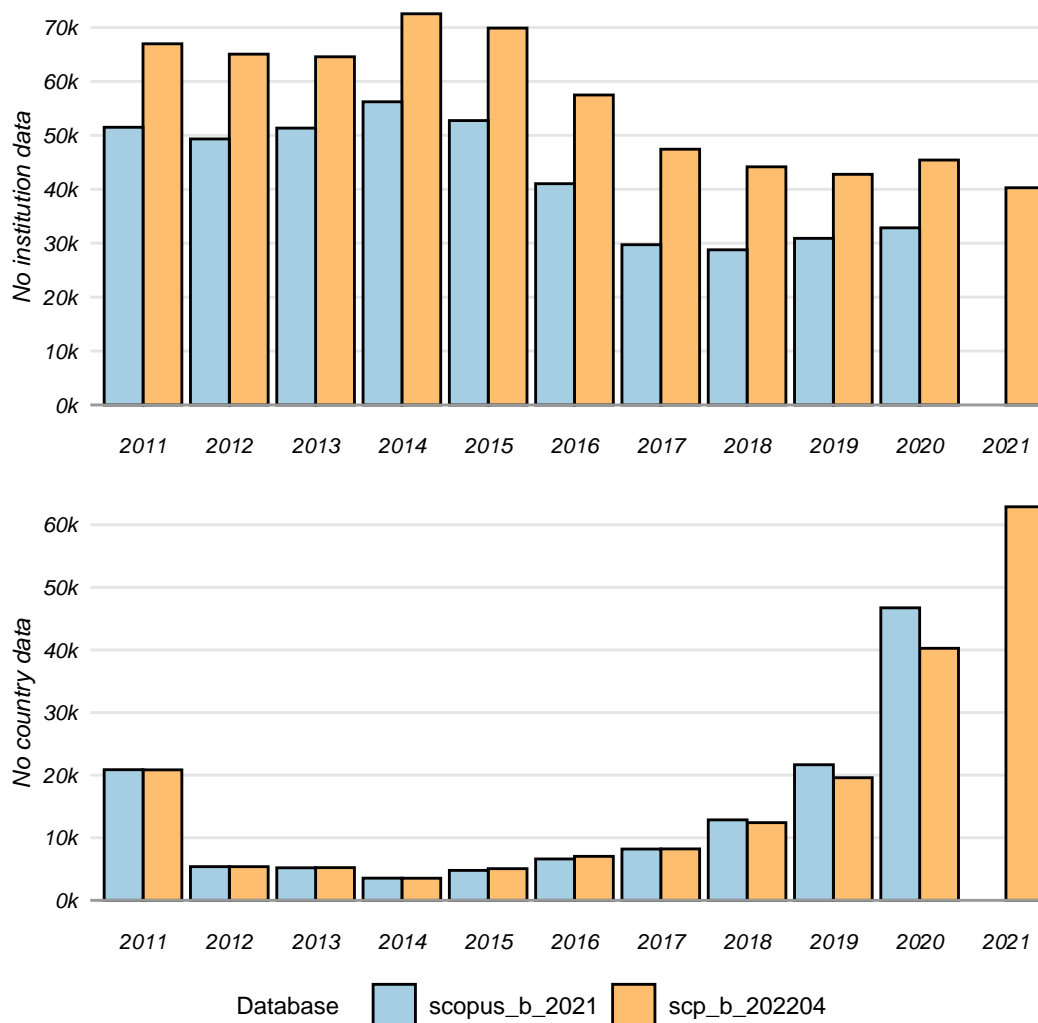


Figure 16: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

## German institutions: German publications missing from KB institution coding

In Figure 17 we show the annual percentage of German publications, i.e. those with a 'DEU' country code in Oracle or where the German indicator is TRUE in PostgreSQL, 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.

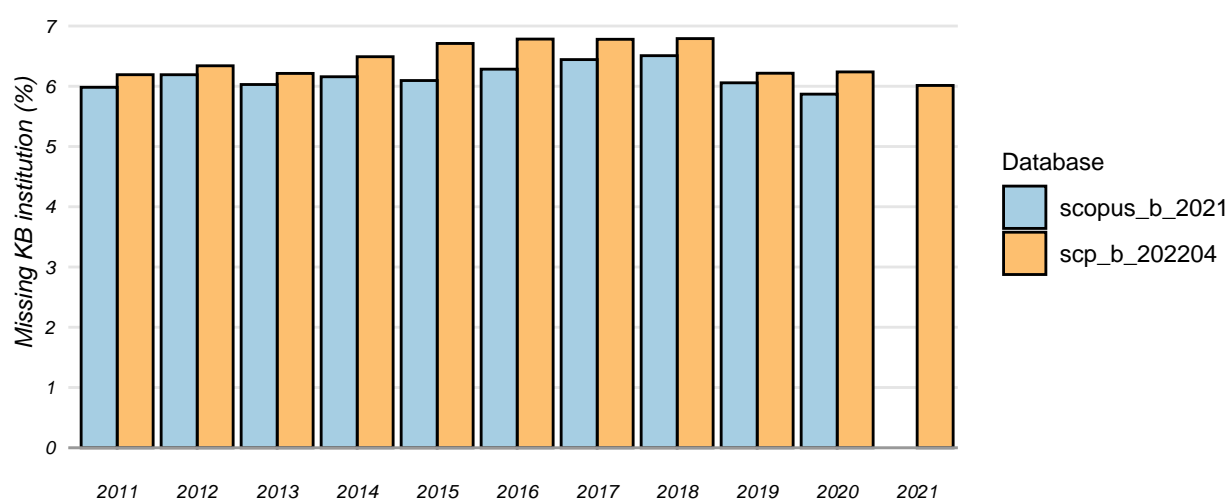


Figure 17: The percentage of German publications in each database that are missing a KB institution.

## 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 5 presents potentially new institutions – these had no publications in 2020 in the scopus\_b\_2021 database but more than five publications in 2021 in the scp\_b\_202204 database. Conversely, Table 6 shows the institutions that had at least five publications in 2020 in the scopus\_b\_2021 database but no publications recorded in 2021 in the scp\_b\_202204 database. We also highlight in Tables 7 and 8 the larger institutions (with at least 20 publications) that had a change in publication counts of more than 40% between 2020 and 2021 in the scopus\_b\_2021 and scp\_b\_202204 databases.

Table 5: Institutions with more than 5 publications in 2021 in scp\_b\_202204 that had no publications in 2020 in scopus\_b\_2021.

Inst ID	Name	Previous pubs	Current pubs
5617	Max-Planck-Institut für Multidisziplinär	0	433

5616	Leibniz-Institut zur Analyse des Biodive	0	226
5470	Max-Planck-Institut für Verhaltensbiolog	0	175
5628	MSH Medical School Hamburg – University	0	174
4766	Helmholtz-Institut Mainz (HIM)	0	172
4700	Helmholtz-Institut Ulm für elektrochemis	0	168
5610	Helmholtz-Institut Erlangen-Nürnberg für	0	153
5612	Helmholtz-Institut Münster (HI MS)	0	151
5641	Hearing4all	0	121
4139	Helmholtz-Institut für Pharmazeutische F	0	112
5660	Heidelberger Institut für Radioonkologie	0	80
5652	MSB Medical School Berlin – Hochschule f	0	79
5662	Pettenkofer School of Public Health Münc	0	79
5674	BMW GROUP	0	77
4765	Helmholtz-Institut Freiberg für Ressourc	0	76
5679	German Institute of Development and Sust	0	68
5480	Leibniz-Institut für Resilienzforschung	0	65
5651	OncoRay – National Center for Radiation	0	62
4181	Nationalpark Bayerischer Wald	0	59
5614	Helmholtz-Institut Würzburg für RNA-basi	0	55
5634	Einstein Center for Neurosciences	0	55
5595	DBFZ Deutsches Biomasseforschungszentrum	0	53
5676	Springer Medizin Verlag GmbH	0	44
5661	MVZ CCB Frankfurt und Main-Taunus GbR	0	41
5615	Fraunhofer Cluster of Excellence Immune-	0	30
5538	Kühne Logistics University – Wissenschaf	0	29
5668	Max-Planck-Zentrum für Physik und Medizi	0	27
4764	Helmholtz International Center for FAIR	0	26
5609	Deutsche Elektronen-Synchrotron DESY	0	26
5643	vivo international e.V.	0	18
4192	Institut für Transfusionsmedizin und Imm	0	17
5671	Hochschule Hamm-Lippstadt	0	16
894	Niedersächsisches Institut für historisc	0	14
5588	CureVac AG	0	14
5636	Studienpraxis Urologie	0	14
5642	Plansee Composite Materials GmbH	0	14
5620	Sigmund Freud PrivatUniversität Berlin	0	13
5631	JCMwave GmbH	0	13
5472	Leibniz-Institut für Finanzmarktforschun	0	12
5619	Alanus Hochschule für Kunst und Gesellsc	0	12
5657	iOMEDICO AG	0	12
5650	MVZ Labor Krone GbR	0	11
412	Kath. Marienkrankenhaus gGmbH	0	8
5590	IQVIA Commercial GmbH & Co. OHG	0	8

5677	Deutsche Gesellschaft für Ortung und Nav	0	8
919	Landesamt für Bergbau, Energie und Geolo	0	7
5630	Deutsche Stiftung Organtransplantation (	0	7
5635	Leberstiftungs-GmbH Deutschland	0	7
5638	mediStatistica	0	7
5639	Pirche AG	0	7
5680	Intel Deutschland GmbH	0	7
5552	Fraunhofer-Einrichtung für Individualisi	0	6
5646	DOG – Deutsche Ophthalmologische Gesells	0	6
5658	Infektiologikum	0	6

Table 6: Institutions with no publications in 2021 in scp\_b\_202204 that had more than 5 publications in 2020 in scopus\_b\_2021.

Inst ID	Name	Previous pubs	Current pubs
3987	Restkategorie Deutschland (keine Zuordnu	862	0
1030	Max-Planck-Institut für biophysikalische	332	0
5	Zoologisches Forschungsmuseum Alexander	150	0
1073	Max-Planck-Institut für experimentelle M	121	0
526	Klinikum Bamberg	18	0
5209	Deutsches Zentrum für Hochschul- und Wis	17	0
5180	KIST Europe Forschungsgesellschaft mbH	16	0
245	Europäische Atomgemeinschaft	12	0
259	Marien-Hospital Wesel gGmbH	10	0
907	Landesanstalt für Landwirtschaft und Gar	10	0
559	Hochschule für angewandte Wissenschaften	8	0
3975	Klinikum St. Elisabeth Straubing GmbH	8	0
4125	Deutsches Forschungsinstitut für öffentl	8	0
4602	Hans-Bredow-Institut für Medienforschung	8	0
532	AMEOS AG	7	0
407	Klinikum Hanau GmbH	6	0
500	Bezirkskrankenhaus Augsburg Klinik für P	6	0
4689	Energiewirtschaftliches Institut an der	6	0
5548	DERMATOLOGIKUM BERLIN Gemeinschaftspraxi	6	0

Table 7: Institutions with more than 20 publications in 2020 in scopus\_b\_2021 that increased in publication counts by over 40% in 2021 in scp\_b\_202204.

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
1541	Daimler AG	27	72	166.7
1047	Max-Planck-Institut für Mathematik	137	327	138.7

4400	Centrum fur Integrierte Onkologie Koln B	29	69	137.9
5488	Helmholtz-Institut fur Funktionelle Mari	45	104	131.1
4758	Max-Planck-Institut fur empirische Asthe	36	82	127.8
5348	Deutsches Zentrum fur Lungenforschung	280	631	125.4
48	Leibniz-Institut fur Atmospharenphysik e	25	55	120.0
1142	Fraunhofer-Institut fur Produktionstechn	25	49	96.0
572	Hochschule fur Technik, Wirtschaft und M	23	44	91.3
628	FOM Hochschule fur Oekonomie & Managemen	38	72	89.5
732	Frankfurt Institute for Advanced Studies	115	215	87.0
545	Hochschule Worms, University of Applied	21	39	85.7
1134	Fraunhofer-Institut fur Toxikologie und	83	153	84.3
103	Universitat der Bundeswehr Munchen	178	325	82.6
1146	Fraunhofer-Institut fur Produktionstechn	45	81	80.0
5140	Restkategorie Universitatskliniken Munch	216	385	78.2
515	St. Hedwig-Krankenhaus Berlin	21	37	76.2
36	Leibniz-Institut fur okologische Raument	29	51	75.9
652	Hochschule Bochum - University of Applie	36	63	75.0
693	Laser Zentrum Hannover e.V. (LZH)	23	40	73.9
2771	United Nations University, Institut fur	26	45	73.1
646	Hochschule fur angewandte Wissenschaften	21	36	71.4
547	Hochschule Ravensburg-Weingarten	31	53	71.0
1637	Zentrum fur Rhinologie und Allergologie	36	61	69.4
1059	Sammelkategorie International Max Planck	133	224	68.4
5368	Translational Lung Research Center Heide	53	89	67.9
552	Technische Hochschule Wildau (FH)	24	40	66.7
1165	Fraunhofer-Institut fur Chemische Techno	37	61	64.9
1606	Bayer Konzern	247	406	64.4
2862	Universitätsklinikum Schleswig-Holstein	952	1563	64.2
625	Hochschule Furtwangen - Informatik, Tech	83	136	63.9
1202	VOLKSWAGEN AG	41	67	63.4
586	Hochschule Magdeburg-Stendal	49	80	63.3
5478	Leibniz-Institut fur Werkstofforientiert	56	91	62.5
5120	Deutsches Zentrum fur integrative Biodiv	300	478	59.3
1143	Fraunhofer-Institut fur Photonische Mikr	21	33	57.1
1181	Fraunhofer-Institut fur Kurzzeitdynamik,	21	33	57.1
5569	Leipzig Heart Institute GmbH	42	66	57.1
56	Fraunhofer-Institut fur Optronik, System	43	67	55.8
1156	Fraunhofer-Institut fur Integrierte Scha	52	80	53.8
481	CTK Cottbus Carl-Thiem-Klinikum gGmbH	28	43	53.6
791	Sanitätsakademie der Bundeswehr, Ernst-v	71	109	53.5
49	Leibniz-Institut fur Agrarentwicklung in	61	93	52.5
144	Zeppelin Universität - Hochschule zwisch	42	64	52.4
675	WHU - Otto Beisheim School of Management	90	137	52.2

913	Bayerisches Landesamt für Gesundheit und	70	106	51.4
50	Heinrich-Pette-Institut für Experimentel	65	98	50.8
17	Museum für Naturkunde Leibniz-Institut f	173	260	50.3
452	St.-Antonius-Hospital Eschweiler	38	57	50.0
629	Hochschule Flensburg	22	33	50.0
4762	Leibniz-Institut für Bildungsverläufe e.	32	48	50.0
5431	Nationales Centrum für Tumorerkrankungen	65	97	49.2
134	Hochschule für Angewandte Wissenschaften	131	195	48.9
4732	Munich Heart Alliance (MHA)	253	374	47.8
289	Klinikum Saarbrücken gGmbH - Akademische	21	31	47.6
367	Kliniken der Stadt Köln gGmbH	96	141	46.9
627	Hochschule Fresenius	30	44	46.7
5290	Deutsches Zentrum für Infektionsforschun	632	926	46.5
5395	Hahn-Schickard-Gesellschaft für angewand	41	60	46.3
1056	Max-Planck-Institut für Innovation und W	22	32	45.5
33	Leibniz-Institut für Pflanzenbiochemie	86	125	45.3
1145	Fraunhofer-Institut für Produktionsanlag	29	42	44.8
634	Frankfurt University of Applied Sciences	43	62	44.2
554	Fachhochschule Südwestfalen	34	49	44.1
11	Senckenberg Gesellschaft für Naturforsch	125	180	44.0
30	Leibniz-Institut für die Pädagogik der N	80	115	43.8
184	Zentrum für systemische Neurowissenschaf	37	53	43.2
1579	Robert Bosch GmbH	97	138	42.3
74	Bernhard-Nocht-Institut für Tropenmedizi	165	233	41.2
4696	European XFEL GmbH	90	127	41.1
83	Universität Vechta	39	55	41.0
5252	Medizinische Hochschule Brandenburg Theo	197	276	40.1

Table 8: Institutions with more than 20 publications in 2020 in scopus\_b\_2021 that decreased in publication counts by over 40% in 2021 in scp\_b\_202204.

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
382	Kerkhoff Klinik	71	42	-40.8
1325	Novartis Deutschland GmbH	47	27	-42.6
135	Technische Universität Hamburg-Harburg	410	217	-47.1
211	Senckenberg Museum für Naturkunde Gortlit	42	22	-47.6
794	Paul-Ehrlich-Institut - Bundesamt für Se	111	53	-52.3
113	Deutsche Sporthochschule Köln	330	157	-52.4
503	Evangelisches Krankenhaus Bielefeld gGmbH	60	27	-55.0
177	Hertie School of Governance	71	23	-67.6
304	Klinikum Ernst von Bergmann	65	21	-67.7

### 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 18 the median number of authors per discipline in 2020 in both databases, and in the right panel the median number of authors per discipline in 2020 in the scp\_b\_202204 database compared to 2021 in the scp\_b\_202204 database.

While little change is expected to be seen in the left-hand panel of Figure 18 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 18, are shown in Table 9.

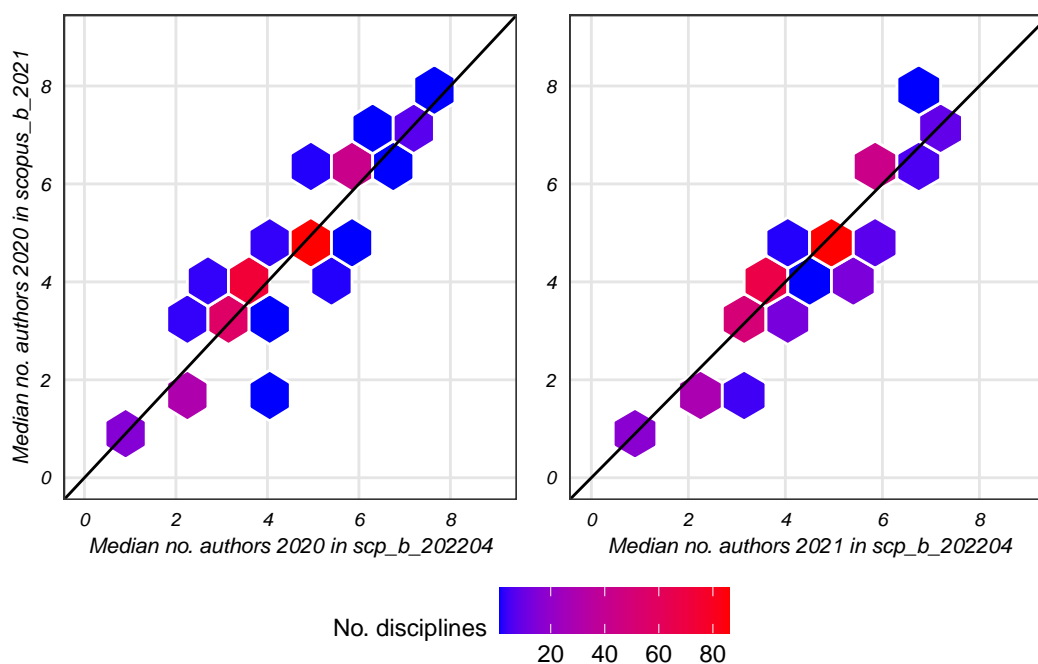


Figure 18: Median number of authors per discipline between databases, where colour denotes the number of disciplines with this combination of median authors.

Table 9: Disciplines where the median number of authors changed by more than 1 between 2020 in scopus\_b\_2021 and 2021 in scp\_b\_202204.

Discipline	Previous median authors	Current median authors	Diff.
------------	-------------------------	------------------------	-------

## 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 non-source items that are not indexed. 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 Scopus' 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 the database provider 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 19 the percentage of references that are source items per discipline in 2020 in both databases, and in the right-hand panel the percentage of references that are source items per discipline in 2020 in the scp\_b\_202204 database compared to 2021 in the scp\_b\_202204 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 19, are shown in Table 10. Longer term trends can be seen in Figure 20 where we present the percentage of reference that are source items per Subject Area over the last ten common years of both databases.

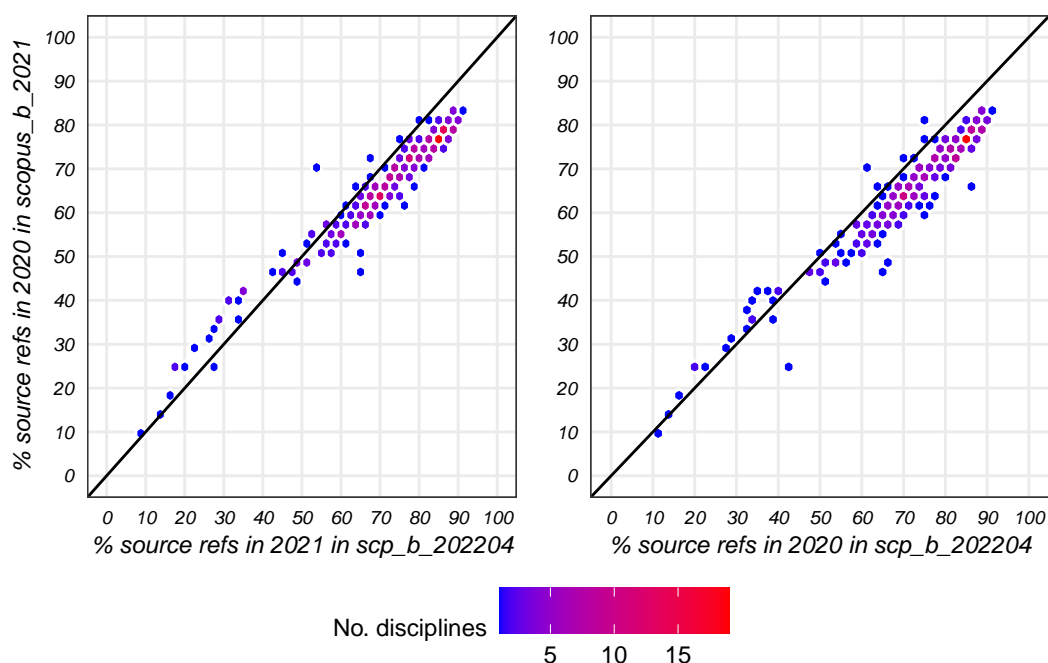


Figure 19: The percentage of cited items that are source items per discipline by database, where colour denotes the number of disciplines with this combination of source references.

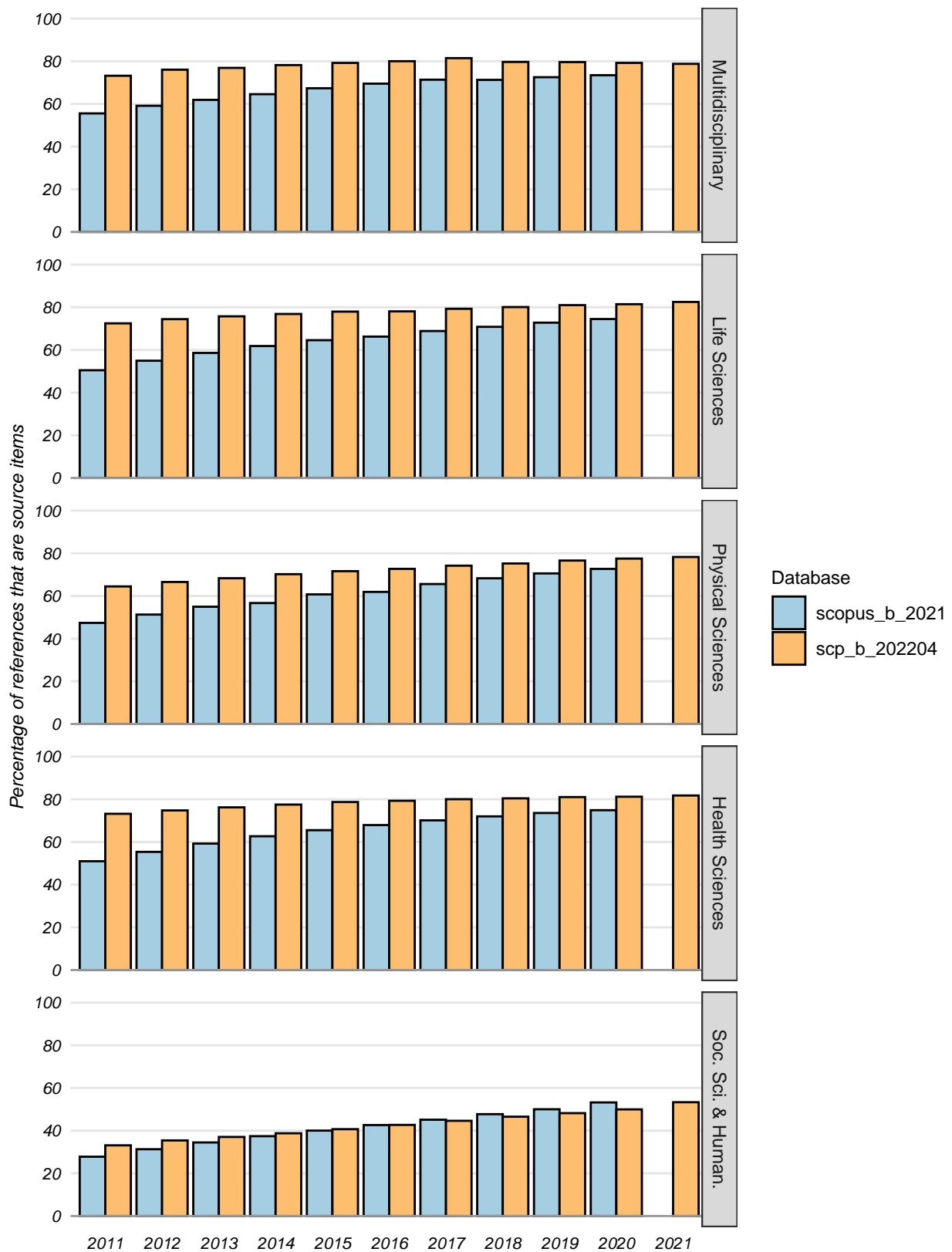


Figure 20: The percentage of references that are source items by Subject Area and database over time.

Table 10: Disciplines where the percentage of indexed references changed by 9 or more percentage points between 2020 in scopus\_b\_2021 and 2021 in scp\_b\_202204.

Discipline	Prvs % source	Crrnt % source	Change
Dental Hygiene	65.9	87.0	21.1
Medical Terminology	24.1	41.6	17.6
Drug Guides	46.8	64.2	17.4
Mathematical Physics	49.6	66.2	16.6
Orthodontics	63.1	77.9	14.8
Veterinary (miscellaneous)	59.1	73.8	14.7
Small Animals	62.9	76.1	13.1
Pharmacology, Toxicology and Pharmaceutics (all)	67.9	80.3	12.4
Equine	63.5	75.9	12.4
LPN and LVN	56.7	68.9	12.2
Reviews and References (medical)	61.7	73.7	12.0
Engineering (miscellaneous)	61.3	72.3	11.0
Endocrine and Autonomic Systems	75.6	86.4	10.9
Periodontics	73.7	84.5	10.8
Endocrinology	76.1	86.7	10.7
Geochemistry and Petrology	58.4	69.1	10.7
Dentistry (miscellaneous)	69.8	80.4	10.6
Management of Technology and Innovation	54.9	65.5	10.5
Cellular and Molecular Neuroscience	78.9	89.4	10.5
Sensory Systems	71.5	81.9	10.5
Neuropsychology and Physiological Psychology	70.7	81.2	10.5
Physiology	75.1	85.5	10.4
Developmental Biology	78.9	89.0	10.1
Finance	52.5	62.5	10.1
Ophthalmology	72.9	82.8	10.0
Otorhinolaryngology	70.2	80.1	10.0
Instrumentation	68.6	78.6	10.0
Paleontology	52.3	62.2	9.9
Physiology (medical)	75.8	85.6	9.9
Behavioral Neuroscience	72.7	82.7	9.9
Food Animals	62.6	72.6	9.9
Plant Science	65.1	74.9	9.8
Inorganic Chemistry	77.3	87.0	9.8
Microbiology	75.5	85.3	9.8
Accounting	50.6	60.4	9.7
Oceanography	59.5	69.2	9.7
Neuroscience (all)	75.5	85.2	9.7
Veterinary (all)	64.5	74.2	9.7

---

Molecular Biology	79.0	88.7	9.6
Cell Biology	81.0	90.4	9.5
Spectroscopy	78.1	87.6	9.5
Computational Mechanics	62.6	72.1	9.5
Neurology	77.4	86.9	9.5
Dentistry (all)	71.2	80.7	9.5
Oral Surgery	72.0	81.6	9.5
Virology	76.4	85.7	9.4
Dermatology	74.6	84.0	9.4
Speech and Hearing	58.8	68.2	9.4
Applied Microbiology and Biotechnology	75.7	85.0	9.3
Anesthesiology and Pain Medicine	74.6	83.9	9.3
Neurology (clinical)	76.6	85.9	9.3
Rehabilitation	69.4	78.6	9.3
Optometry	70.0	79.3	9.3
Aquatic Science	60.3	69.4	9.1
Clinical Biochemistry	79.3	88.4	9.1
Structural Biology	78.0	87.1	9.1
Organic Chemistry	77.0	86.1	9.1
Embryology	72.5	81.5	9.1
Genetics (clinical)	79.5	88.6	9.1
Developmental Neuroscience	75.0	84.2	9.1
Physical and Theoretical Chemistry	76.4	85.4	9.0
Medicine (all)	73.3	82.2	9.0
Reproductive Medicine	73.7	82.7	9.0
Neuroscience (miscellaneous)	73.5	82.5	9.0
Biological Psychiatry	76.5	85.5	9.0
Toxicology	73.0	82.0	9.0
Decision Sciences (miscellaneous)	69.6	60.2	-9.5

---

## References

- [1] 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.