

---

**Quality assurance at the macro level: Comparing the current and  
previous WoS 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 Research 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 . . . . .	13
Disciplines: Changes in discipline classification . . . . .	15
Disciplines: Changes in articles and reviews by discipline . . . . .	16
Disciplines: Percentage of publications not assigned to a discipline . . . . .	17
Metadata: Changes in pubyear, doctype, pubtype and items removed . . . . .	17
Metadata: Publications from each index . . . . .	19
Metadata: Missing metadata variables . . . . .	20
Institution and country data: Number of articles and reviews with missing data . . . . .	21
Author-institution links: Percentage complete by Research Area and discipline . . . . .	22
German institutions: German publications missing from KB institution coding . . . . .	24
German institutions: Changes in whole counts of articles and reviews . . . . .	24
Authors: Median number of authors by Research Area and discipline . . . . .	28
Source items: Percentage by Research 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 wos\_b\_2021 and wos\_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 wos\_b\_2021 and wos\_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 wos\_b\_2021 database against 2019 in the wos\_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. Unless otherwise stated, we examine content indexed in the Science Citation 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 [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 the `sc_traditional` or Research Areas (RA) classification. `Sc_traditional` is a fine-grained classification that allows changes in specific disciplines to be analysed. However, as 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 RA classification. The RA consists of five broad groups: Life Sciences, Physical Sciences, Technology, Social Sciences, Arts and Humanities, and Multidisciplinary. These groups are usually mapped from the `sc_extended` disciplines provided by Clarivate Analytics. However, as the PostgreSQL version of the databases contains only the `sc_traditional` classification, the mapping has been conducted based on this classification.

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 Research 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 version, 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 RA.

Changes in publication counts over time may reflect changes made by countries, the database provider, and/or administrative decisions. For example, it is expected that the `wos_b_202204` database contains a greater number of publications for the most recent years than the `wos_b_2021` database due to the continued indexing of items by Clarivate Analytics 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 WoS' 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 WoS-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.

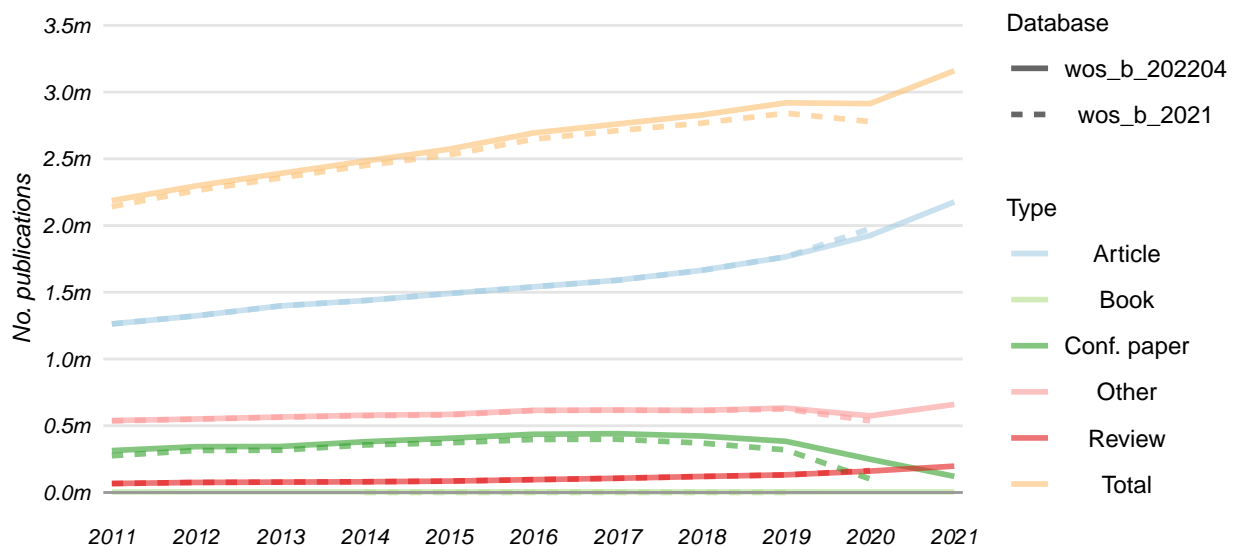


Figure 1: Number of documents in each database over time by type.

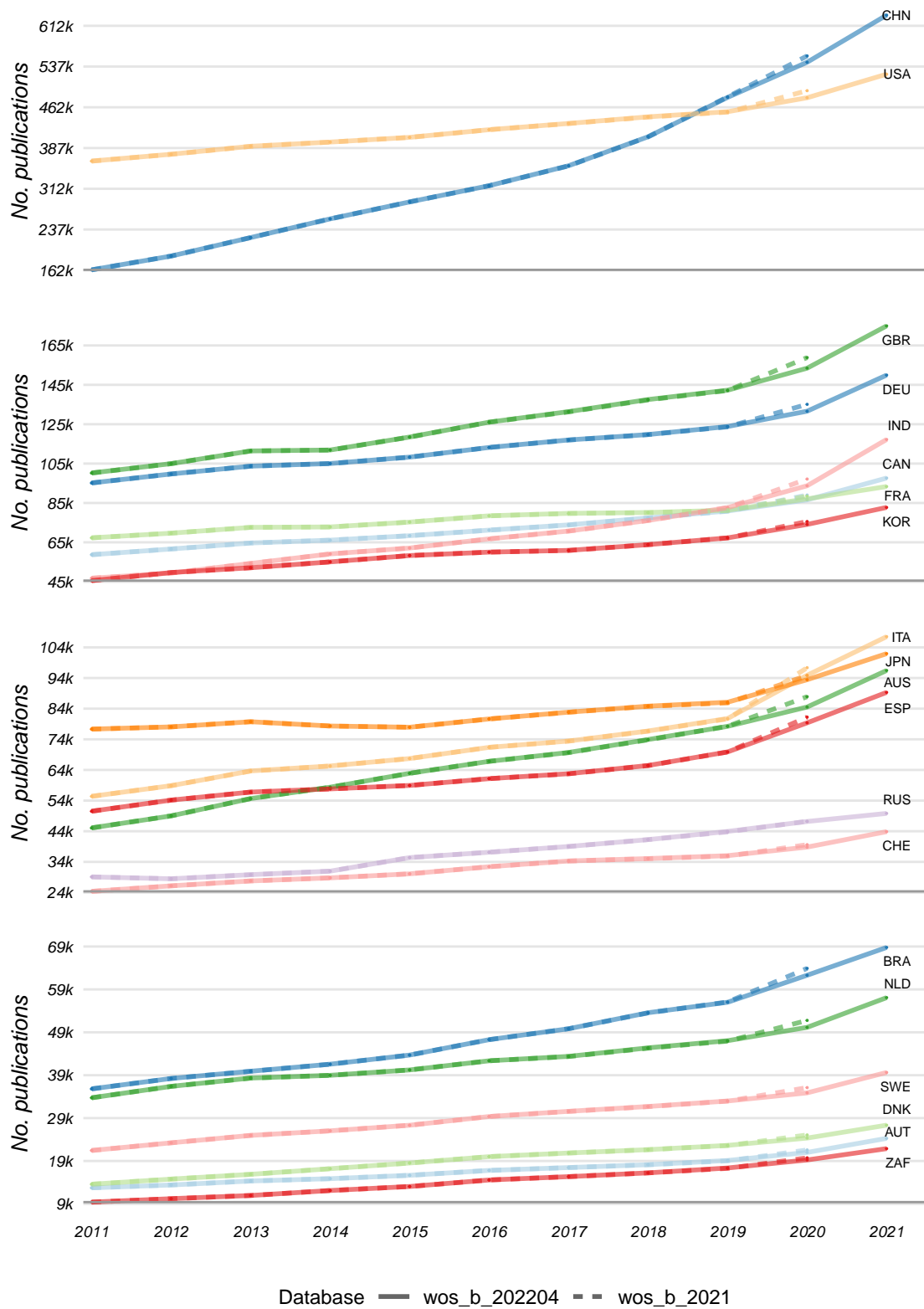


Figure 2: Whole counts of articles and reviews by country and database over time. Please note the panels' different axes.

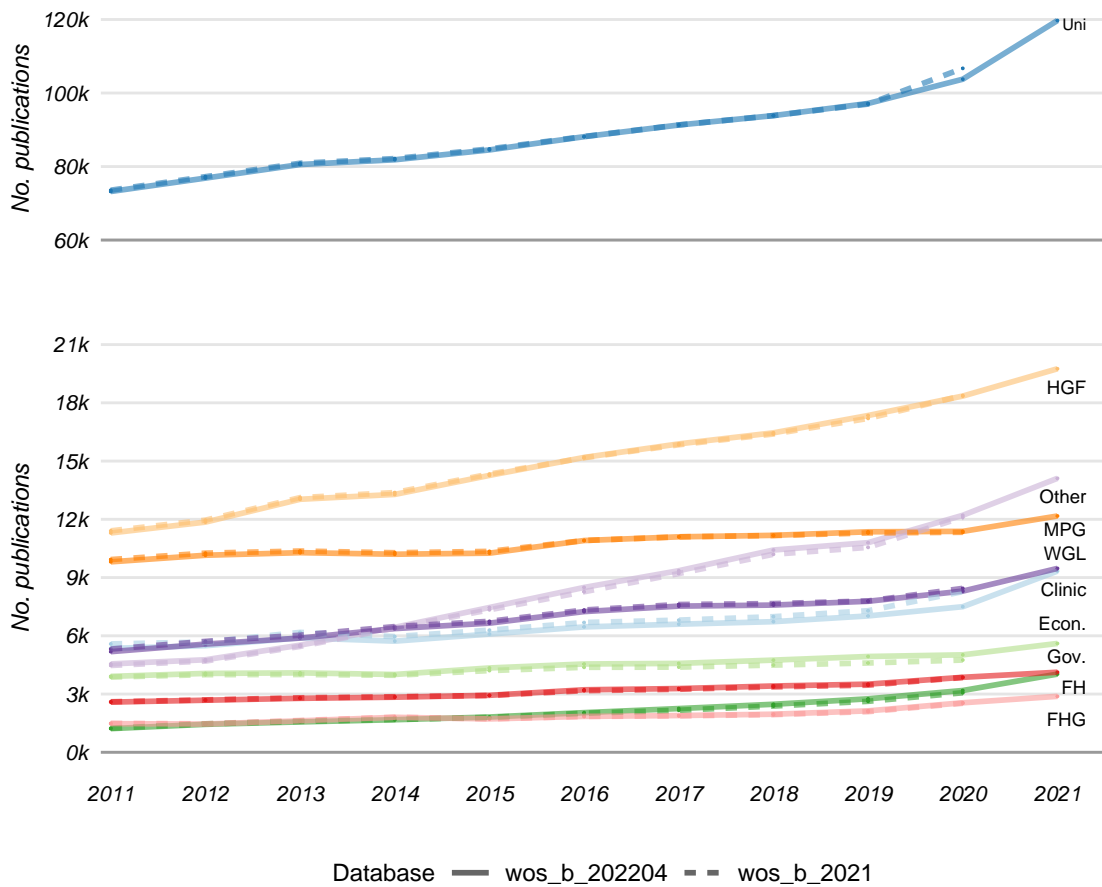


Figure 3: Whole counts of articles and reviews by German sector and database over time. Please note the panels' different axes.

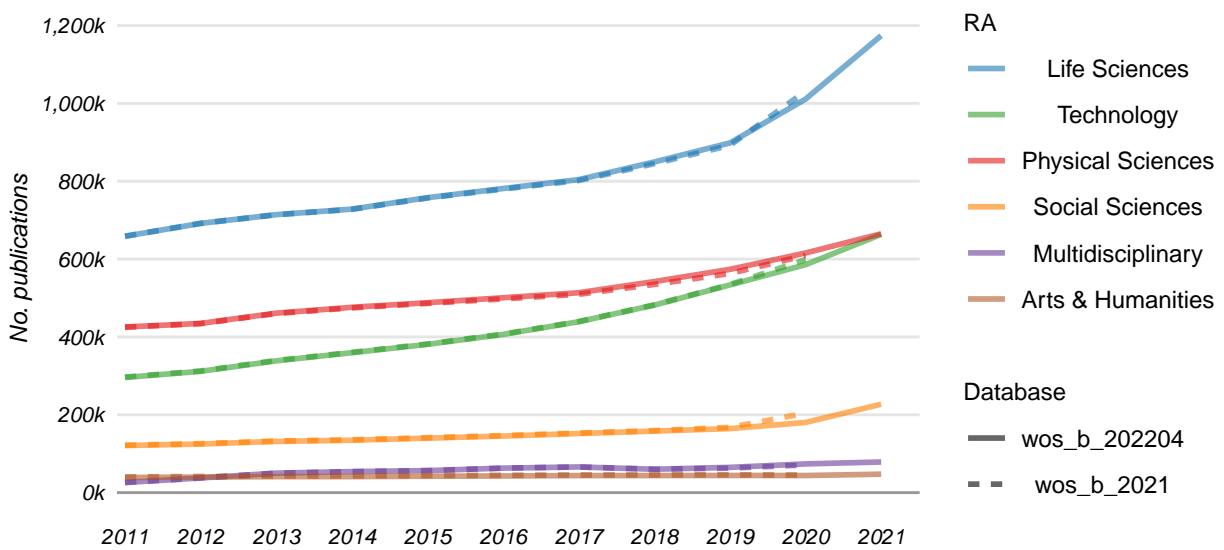


Figure 4: Whole counts of articles and reviews by RA and database over time.

## 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 RA.

Changes in the journals indexed were identified by matching the titles of all journals indexed in 2020 in the *wos\_b\_2021* database to those with 2021 content in the *wos\_b\_202204* database. Titles were used as all journals have titles recorded, while some journals are missing ISSNs. Titles in *wos\_b\_2021* but not in *wos\_b\_202204* were considered removed, while titles in *wos\_b\_202204* but not in *wos\_b\_2021* were considered added. In total, 325 journals were added and 248 were removed. These data may include a small number of journals that changed titles. Some double-counting of journals between RAs may also occur when a journal maps to two or more RAs.

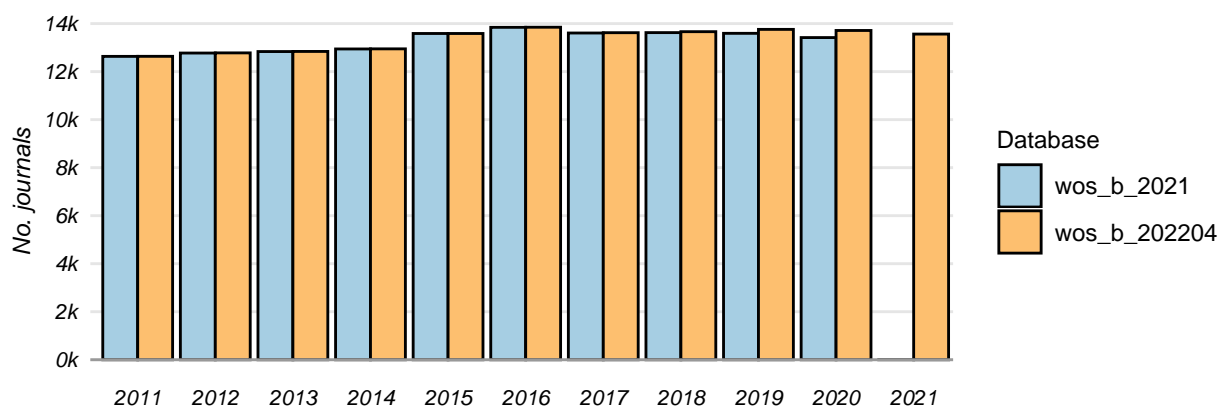


Figure 5: The number of journals indexed in each database over time.

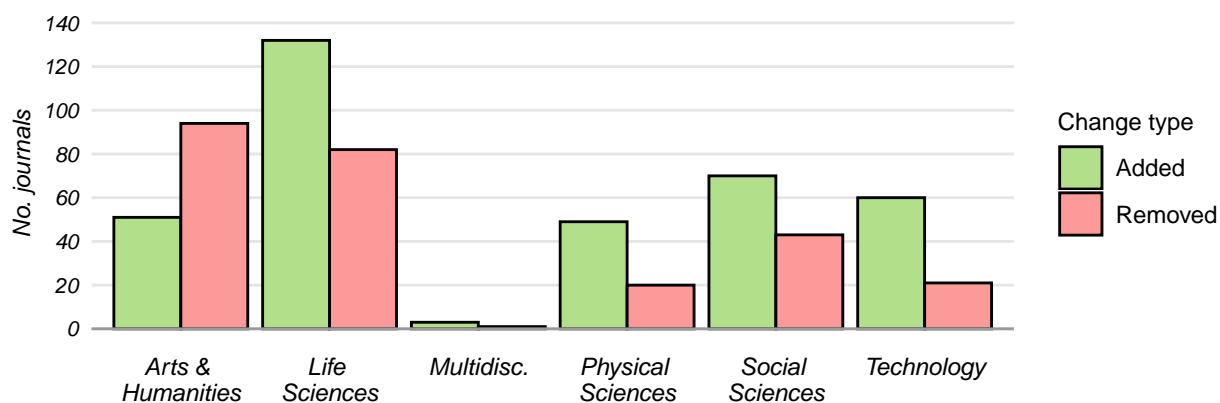


Figure 6: The number of journals added or removed between 2020 in *wos\_b\_2021* and 2021 in *wos\_b\_202204* by RA.

### 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.

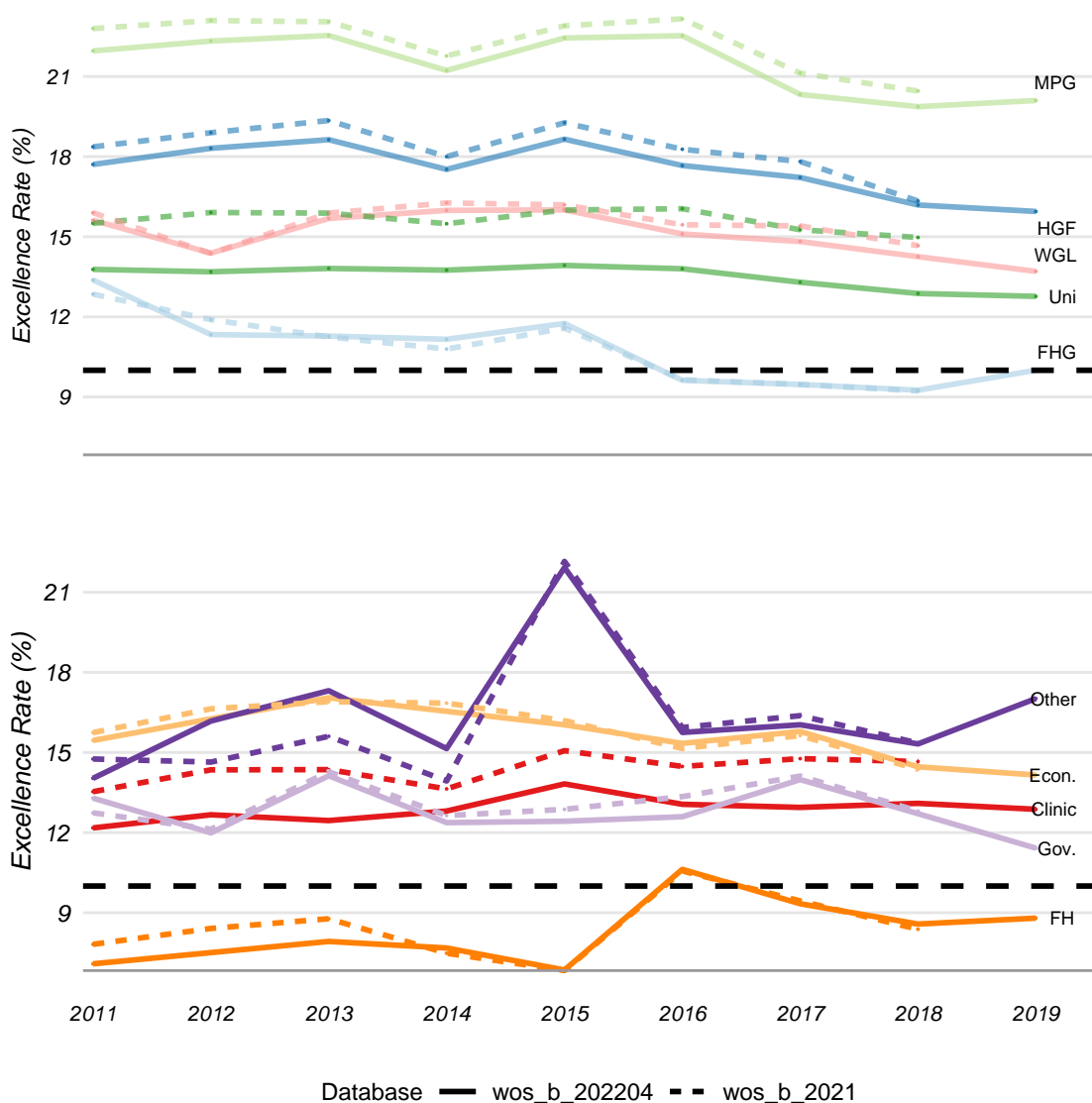


Figure 7: ERs, based on whole counts, by German sector and database over time. The black line is the expected 10% threshold.

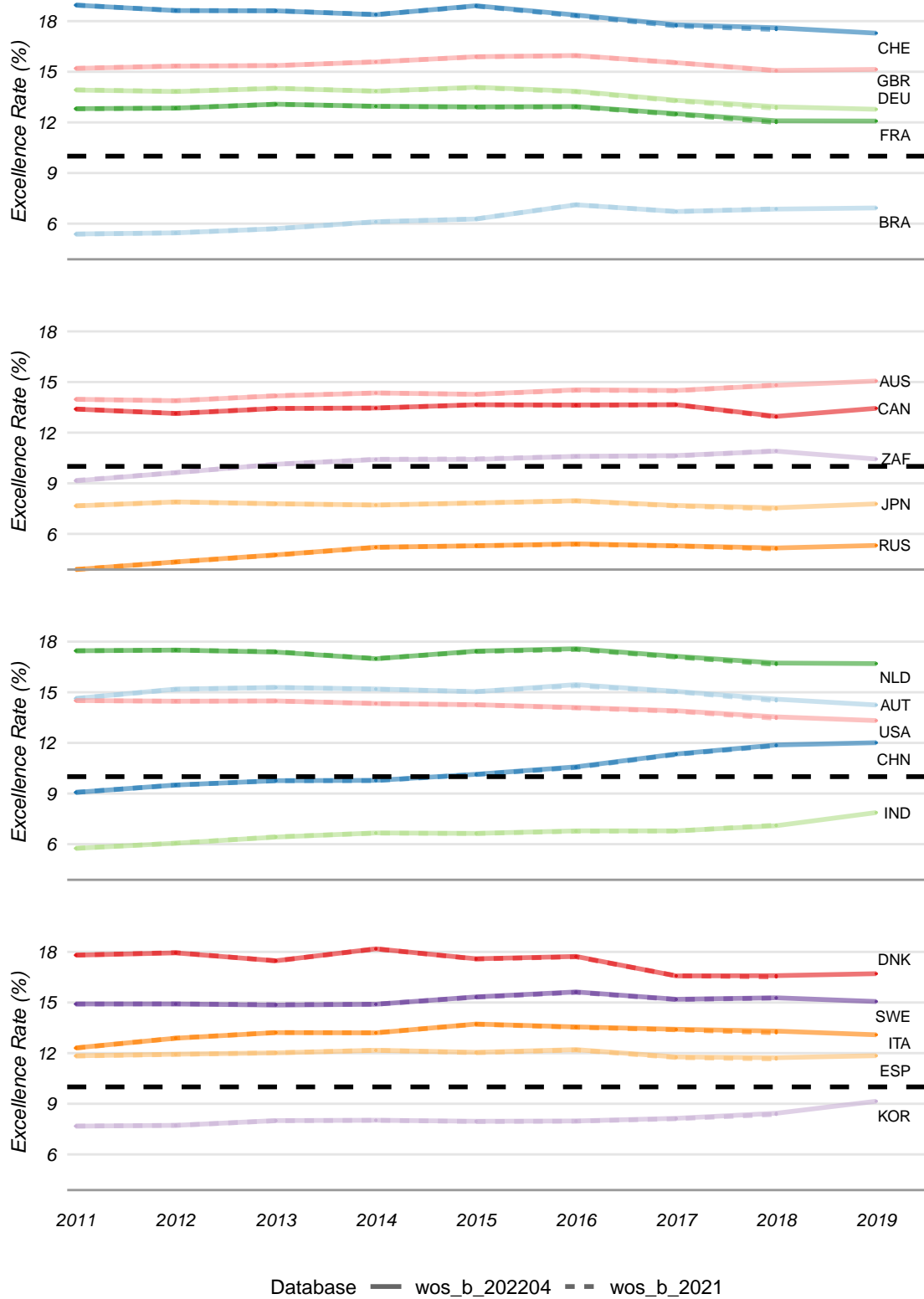


Figure 8: ERs, based on whole counts, by selected country and database over time. The black line is the expected 10% threshold.

## 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.

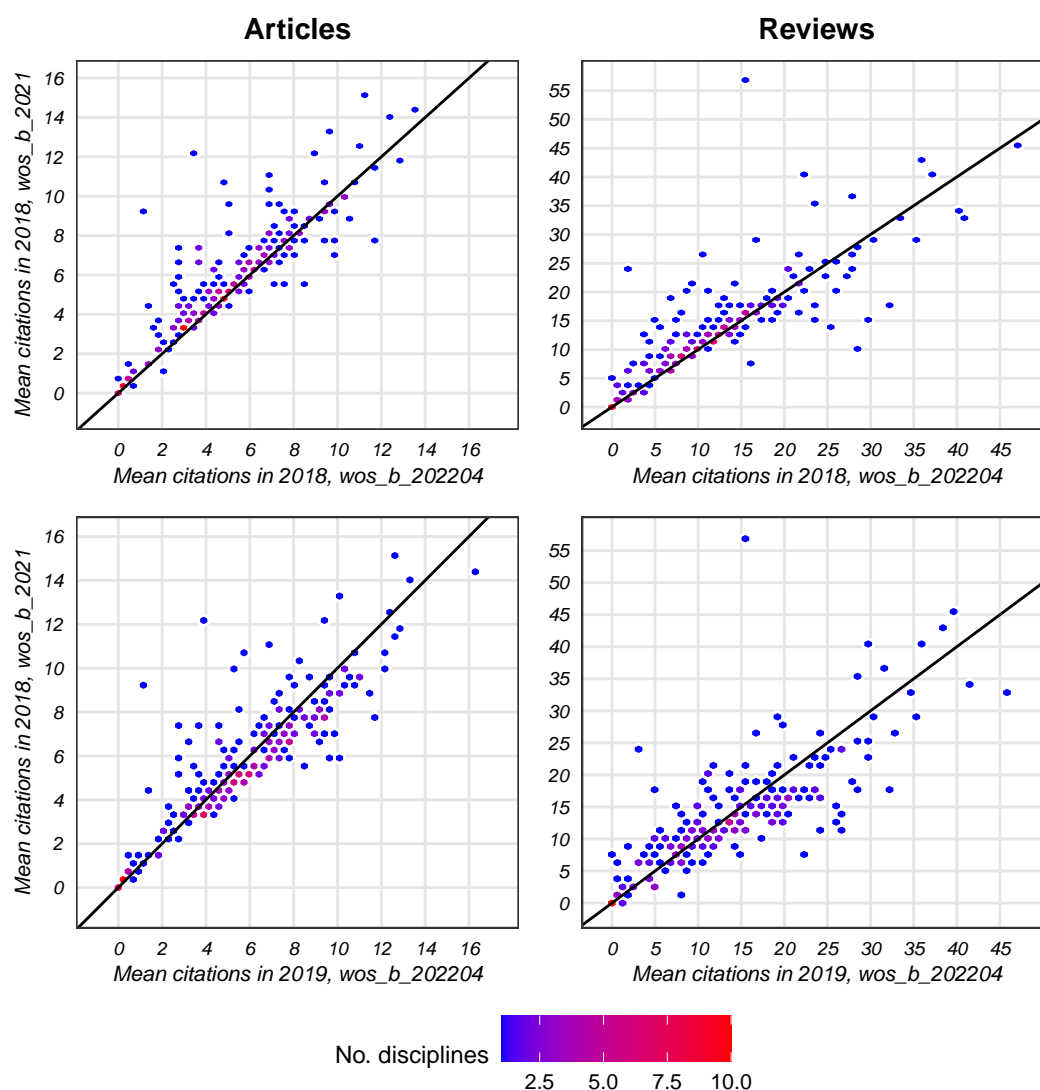


Figure 9: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

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 wos\_b\_2021 and 2019 in wos\_b\_202204, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. crnt pubs.	Perc. diff.
Ergonomics	5.7	10.0	548	75.8
Quantum Science & Technology	5.7	9.7	726	69.0
Transportation Science & Technology	7.8	11.9	457	51.8
Urban Studies	5.7	8.5	482	49.6
Soil Science	7.0	9.7	2453	39.4
Cultural Studies	1.3	1.8	1244	36.0
Robotics	7.1	9.6	1547	35.6
Business	6.8	9.0	8589	33.5
Physics, Particles & Fields	8.7	11.5	3705	32.0
Regional & Urban Planning	6.5	4.4	210	-32.0
Medical Ethics	3.6	2.4	42	-32.8
Telecommunications	8.2	5.5	2067	-32.9
Materials Science, Textiles	5.5	3.5	1781	-36.4
Materials Science, Multidisciplinary	11.0	7.0	47278	-36.8
Operations Research & Management Science	7.2	4.5	1920	-37.2
Crystallography	5.1	2.9	1984	-42.0
Computer Science, Cybernetics	10.6	5.9	994	-44.7
Physics, Applied	9.8	5.3	10671	-45.8
Instruments & Instrumentation	7.2	3.7	5399	-48.9
Metallurgy & Metallurgical Engineering	6.7	3.1	4758	-53.8
Mineralogy	6.0	2.6	364	-56.6
Physics, Atomic, Molecular & Chemical	7.2	2.8	723	-60.7
Tropical Medicine	4.6	1.5	155	-67.8
Physics, Condensed Matter	12.3	3.9	3418	-68.4
Imaging Science & Photographic Technology	9.2	1.1	97	-88.5

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 wos\_b\_2021 and 2019 in wos\_b\_202204, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. crnt pubs.	Perc. diff.
Film, Radio, Television	0.0	1.0	4	Inf
Literature, Romance	0.0	0.3	3	Inf
Archaeology	1.5	8.0	31	440.6
Urban Studies	7.5	21.8	12	190.0
Mining & Mineral Processing	10.9	26.2	9	141.6
Regional & Urban Planning	10.8	23.8	6	120.2
Materials Science, Paper & Wood	12.6	25.7	44	103.5
Hospitality, Leisure, Sport & Tourism	14.2	26.7	161	88.3
Information Science & Library Science	9.7	17.8	97	83.6
Business, Finance	4.7	8.5	70	80.6
Soil Science	18.1	32.6	49	80.0
Transportation Science & Technology	14.6	26.2	15	80.0
Agricultural Economics & Policy	8.3	14.8	29	78.2
Anthropology	2.7	4.8	100	77.7
Computer Science, Cybernetics	8.3	14.7	23	76.7
Ethnic Studies	2.9	5.2	12	76.4
Computer Science, Artificial Intelligence	17.3	28.2	310	63.1
Engineering, Marine	11.9	9.4	59	-21.1
Chemistry, Organic	13.4	10.6	842	-21.4
Neuroimaging	19.5	15.2	42	-22.0
Microscopy	12.1	9.3	15	-23.1
Mathematics, Interdisciplinary Applications	12.6	9.6	53	-23.5
Medical Informatics	15.9	12.0	54	-24.6
Anatomy & Morphology	10.2	7.7	176	-24.6
Mineralogy	12.7	9.5	6	-25.3
Communication	7.4	5.5	73	-25.9
Meteorology & Atmospheric Sciences	18.4	13.4	53	-26.9
Physics, Applied	40.6	29.4	244	-27.7
Transplantation	8.6	6.2	159	-27.8
Psychology, Social	12.9	9.3	64	-28.1
Engineering, Mechanical	21.2	15.2	221	-28.2
Psychology, Experimental	14.7	10.4	65	-29.2
Astronomy & Astrophysics	28.3	19.6	265	-30.8
Engineering, Chemical	28.8	19.8	511	-31.3
Remote Sensing	20.3	14.0	24	-31.3
Nuclear Science & Technology	10.1	6.8	59	-33.0
Zoology	6.4	4.2	303	-33.9
Ergonomics	10.1	6.7	20	-34.1

---

Geology	12.0	7.9	70	-34.4
Materials Science, Composites	26.9	17.3	60	-35.8
Crystallography	15.3	9.6	77	-37.3
Geography	9.2	5.7	47	-37.5
Paleontology	6.2	3.9	16	-37.7
Physics, Atomic, Molecular & Chemical	18.5	11.1	10	-39.8
Social Issues	6.2	3.7	26	-40.9
Public Administration	13.9	8.1	16	-41.8
Mathematical & Computational Biology	18.5	10.7	48	-42.3
Water Resources	19.8	11.1	50	-43.9
Operations Research & Management Science	21.3	11.8	38	-44.8
Evolutionary Biology	19.9	10.9	41	-45.3
Materials Science, Textiles	9.6	5.2	67	-45.8
Social Work	8.5	4.5	52	-47.3
Physics, Mathematical	8.2	3.9	49	-52.3
Engineering, Ocean	14.7	7.0	4	-52.5
Mathematics, Applied	9.7	4.5	73	-53.8
Materials Science, Characterization & Testing	11.9	5.3	23	-56.0
Instruments & Instrumentation	17.9	5.1	393	-71.6
Physics, Condensed Matter	56.4	15.7	58	-72.2

---

## 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.

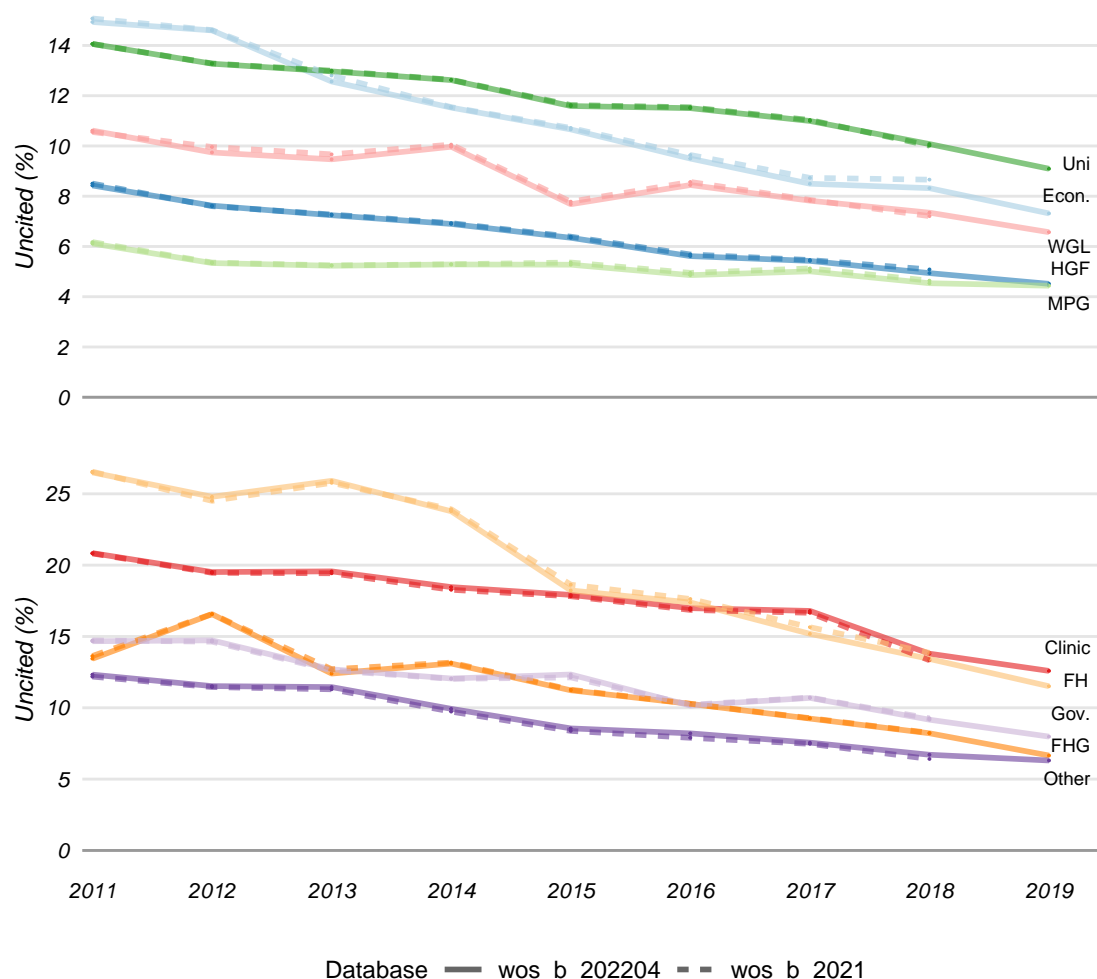


Figure 10: The percentage of uncited publications in each database over time by German sector.

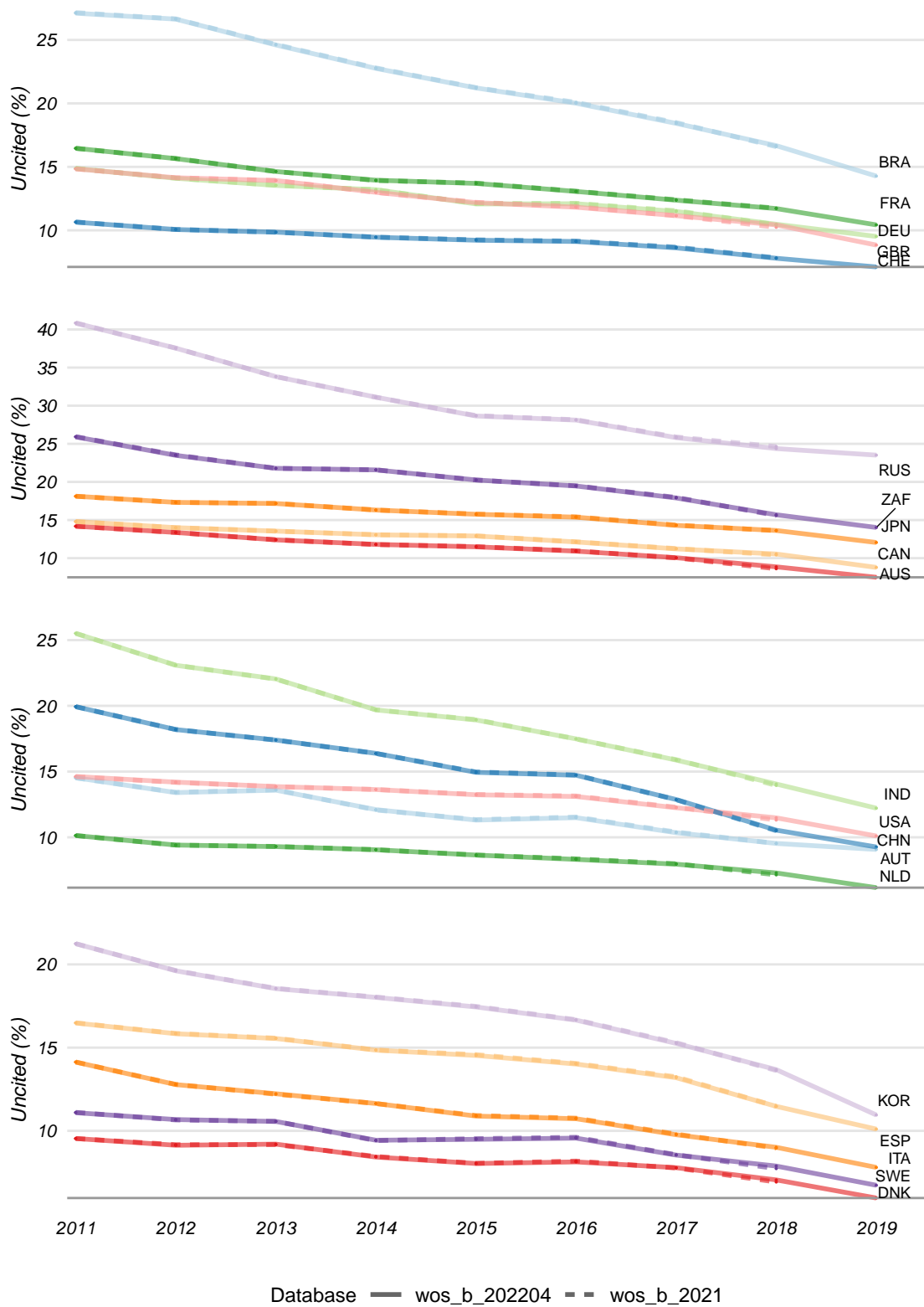


Figure 11: The percentage of uncited publications in each database over time by selected countries.

## Disciplines: Changes in discipline classification

This section highlights any changes that have been made to WoS' sc\_traditional classification in Table 3. This could include splits, aggregations or removals of a discipline, or the inclusion of a new discipline to reflect new and emerging topics. We identify changes in the classification structure by comparing the number of articles and reviews attributed to each discipline in the latest years of each database and selecting those disciplines where the number was zero in one year but not in the other. Disciplines with no prior publications but some in the current year suggest the discipline may have been recently added, while the opposite suggests the discipline may have been removed or merged. Changes may also reflect changes in spelling or punctuation of the discipline name. Any changes should be checked with WoS' published classification structure.

Table 3: Changes in the sc\_traditional discipline classification structure between the previous and current databases

Classification	Previous pubs	Current pubs
----------------	---------------	--------------

Figure 12 shows the number of publications assigned to specific sc\_traditional disciplines known to have changed in recent versions of the database. The *Planning & Development* subject category also known to have previously been affected by changes was also examined here. No publications in the reference period were assigned to this category in either database version. However, 1,188 publications published during 1986-1994 were assigned to *Planning & Development* in the current PostgreSQL database. Due to the age and small number of publications involved, this issue is unlikely to have any effect on analyses. No publications were assigned to *Planning & Development* in the Oracle database version.

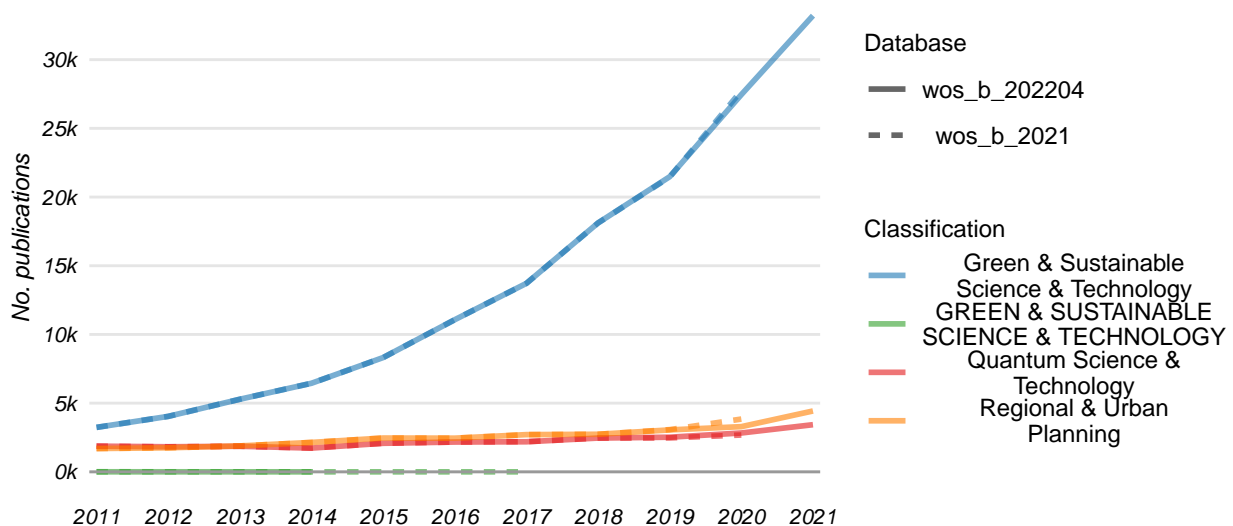


Figure 12: Time-series of sc\_traditional disciplines previously observed to have changed.

## 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 *wos\_b\_2021* and 2021 in *wos\_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 *sc\_traditional* classification. Disciplines previously identified as being new or removed have not been included here.

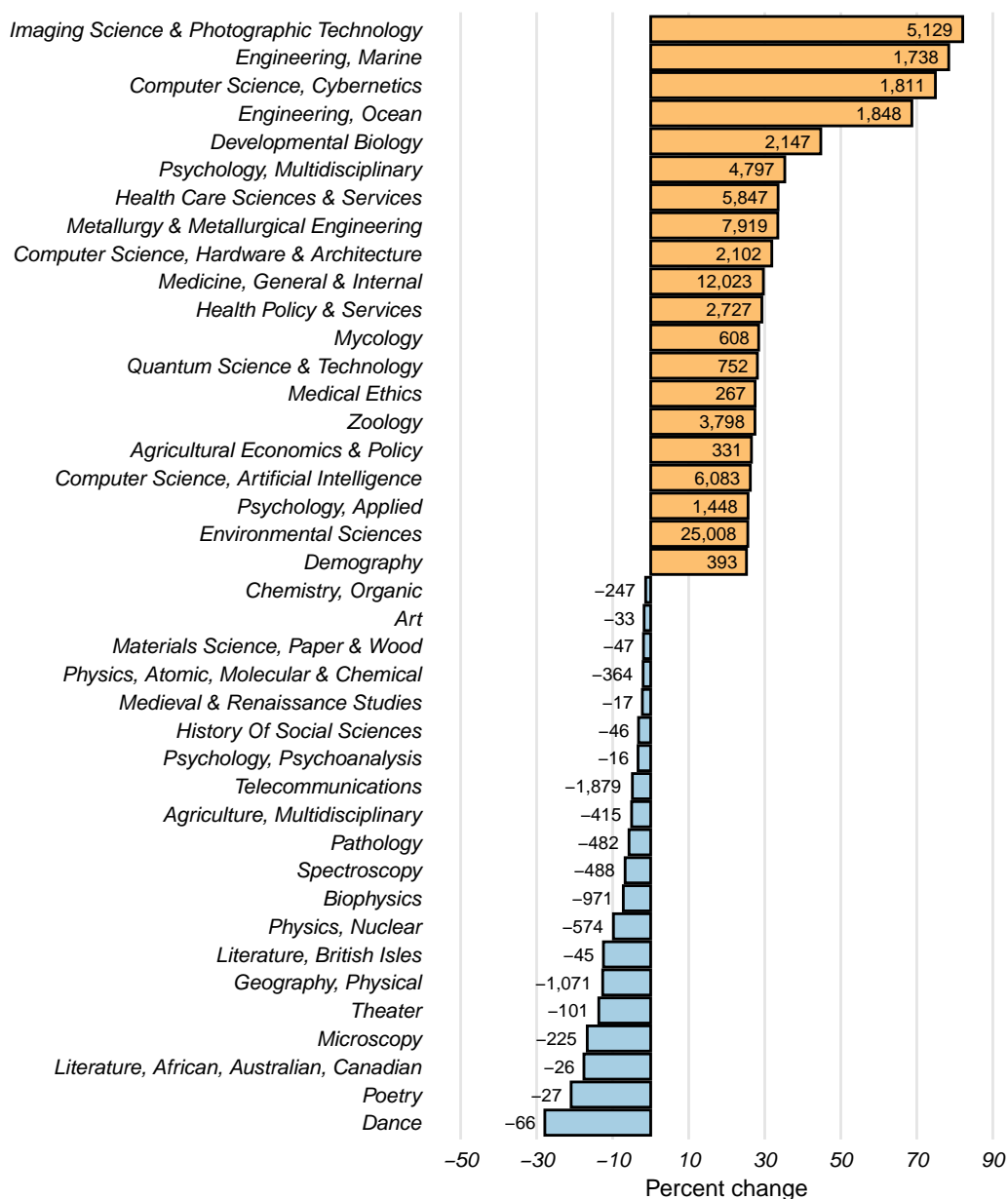


Figure 13: The 40 disciplines with the highest percentage change in publication counts between 2020 in the previous database and 2021 in the current database, with numerical difference in counts.

## 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.

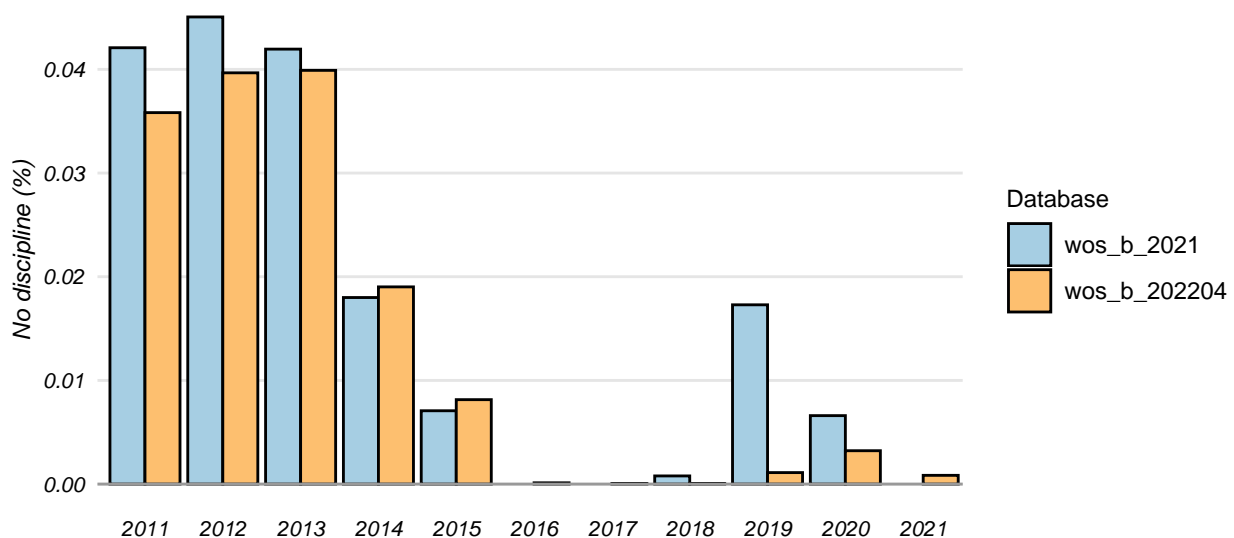


Figure 14: The percentage of publications in each database that do not have a discipline classification.

## 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 wos\_b\_2021 database but not in the wos\_b\_202204 database (removed items), and items present in the wos\_b\_202204 database but not in the wos\_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 wos\_b\_2021 and wos\_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 wos\_b\_2021 and wos\_b\_202204.

Crrnt year	Prvs year	Diff. year	Diff. pubtype	Diff. doctype	Added	Removed
2016	2016	0	161	46354	0	0
2016	2017	190	0	0	0	0
2016	0	0	0	0	4926	0
2017	2017	0	1866	44325	0	0
2017	0	0	0	0	9909	0
2018	2017	91	0	2	0	0
2018	2018	0	1553	0	47148	0
2018	2019	33	0	0	0	0
2018	0	0	0	0	18561	0
2019	2018	26	0	0	0	0
2019	2019	0	471	48663	0	0
2019	2020	252	0	2	0	0
2019	0	0	0	0	44438	0
2020	2017	12	0	0	0	0
2020	2018	10	0	0	0	0
2020	2019	741	0	6	0	0
2020	2020	0	181	54028	0	0
2020	0	0	0	0	176197	0
0	2016	0	0	0	0	366
0	2017	0	0	0	0	891
0	2018	0	0	0	0	375
0	2019	0	0	0	0	10724
0	2020	0	0	0	0	92593

### Metadata: Publications from each index

The WoS database is comprised of several indices. The KB contract with Clarivate Analytics specifies that we receive data from the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (AHCI). The other indices are the Book Citation Index (Science, BSCI), Conference Proceedings Citation Index (Science, ISTP), Conference Proceedings Citation Index (Social Sciences & Humanities, ISSHTP), Current Chemical Reactions (CCR), Emerging Sources Citation Index (ESCI), and Index Chemicus (IC). The inclusion of items from other indices can be problematic as these items may fundamentally differ from those in the three core indices in, for instance, the countries of their authors or publishing journals, which can influence citation-based indicators. As such, we examine in Figure 15 the number of articles and reviews in each index in the `wos_b_2021` and `wos_b_202204` databases between 2011 and 2021. As it is possible in the `wos_b_202204` database for articles/reviews to also be classified as conference papers, these documents may also appear in the ISTP and ISSHTP indices, in addition to the SSCI, SCIE and AHCI.

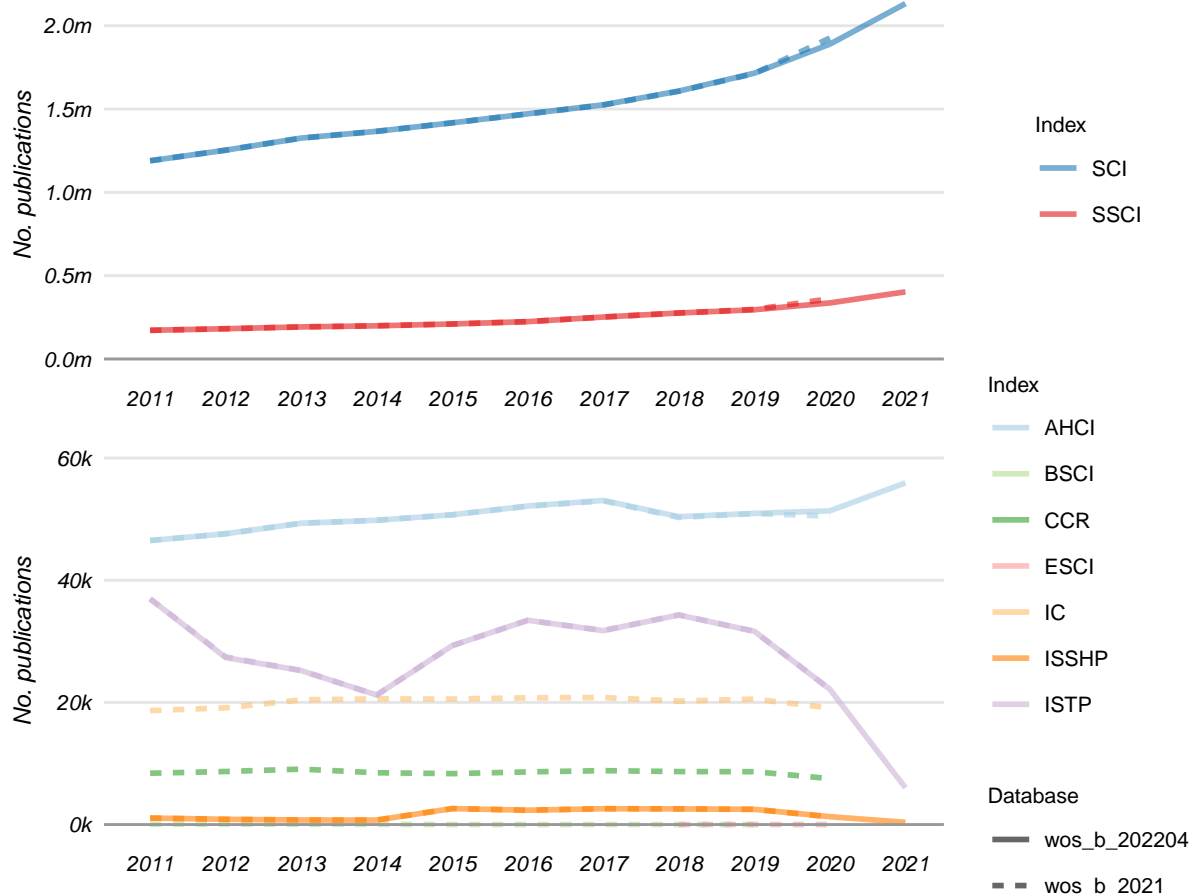


Figure 15: The number of articles and reviews in each WoS index by database over time.

## 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.

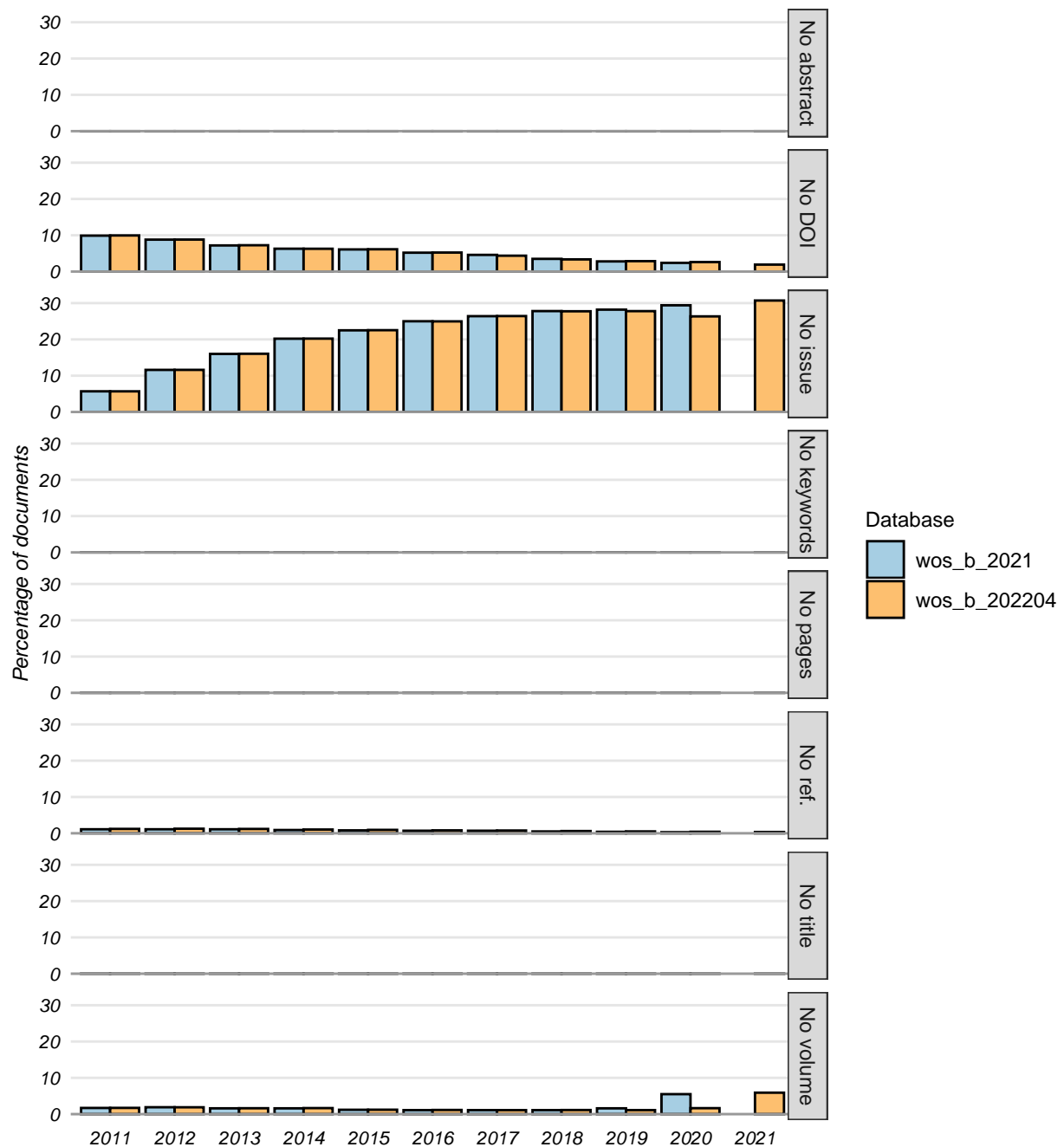


Figure 16: The percentage of items with missing metadata over time by database.

### 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 in `wos_b_2021` and in the `items_affiliations` table in `wos_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 17 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 17. Large disparities between the databases or substantial increases in missing information should be investigated.

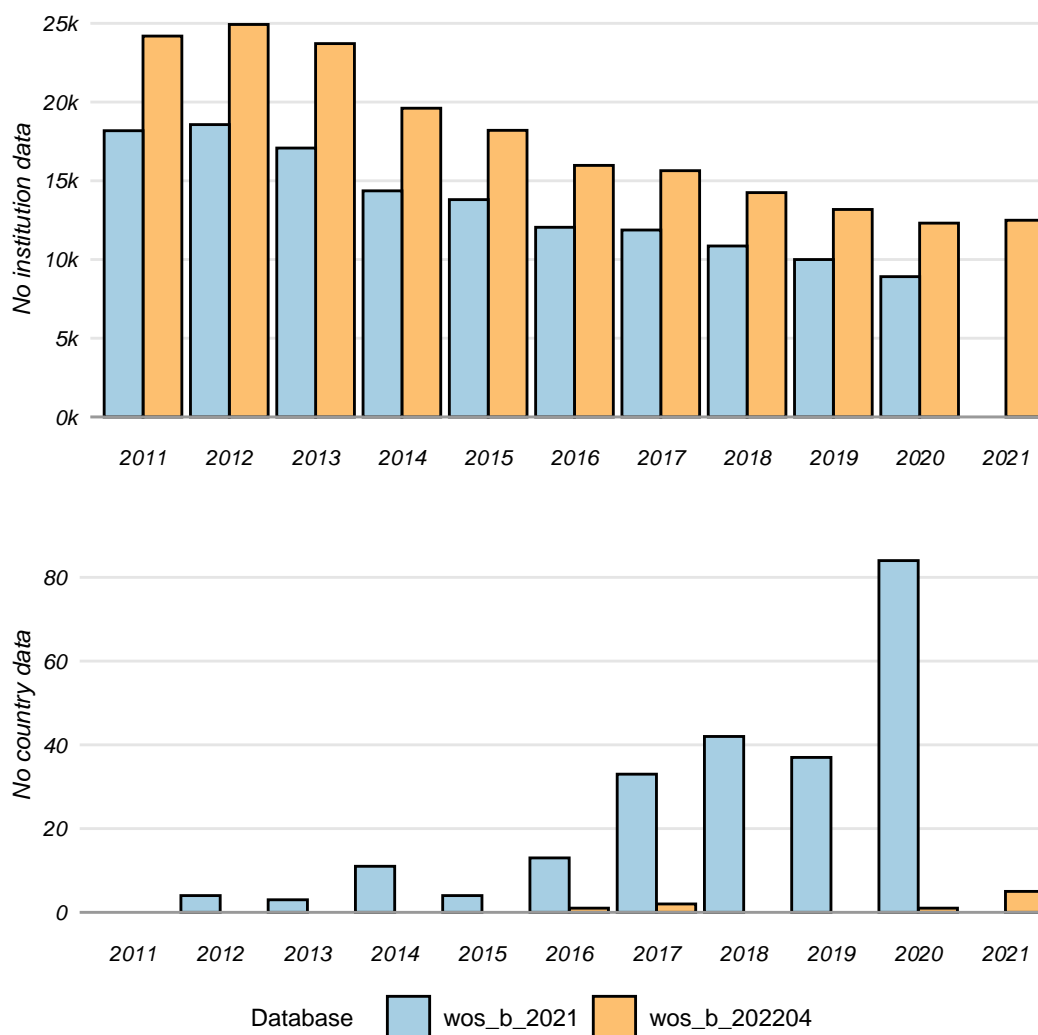


Figure 17: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

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

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

In Figure 18, we see in the left panel the percentage of complete links for 2020 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 2020 in the *wos\_b\_2021*, now compared with the 2021 in the *wos\_b\_202204*, indicating potential changes between the latest year in each database.

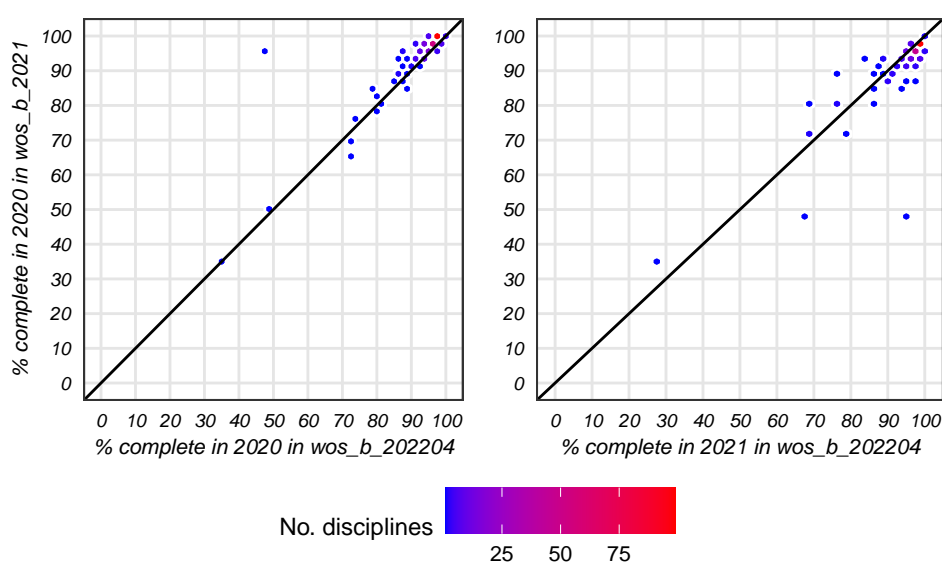


Figure 18: The percentage of complete author-institution links by disciplines.

Table 5 shows the outlying disciplines observable in the right panel of Figure 18 that changed by more than 7 percentage points in the percentage of complete author-institution links.

Table 5: Disciplines that changed by more than 7 percentage points in missing links between 2020 in *wos\_b\_2021* and 2021 in *wos\_b\_202204*.

Discipline	Prvs %	Crrnt %	Prvs no.	Crrnt no.	Change
Psychology, Psychoanalysis	47.9	93.8	239	436	-45.86
Dance	48.5	66.7	123	114	-18.17
Ornithology	87.7	96.7	1034	1143	-9.00
Dentistry, Oral Surgery & Medicine	85.4	94.3	9656	11010	-8.85
Primary Health Care	86.3	95.0	1562	1641	-8.67
Literary Reviews	35.4	27.3	436	343	8.07
Literature, British Isles	93.1	84.2	339	267	8.90
Literature, African, Australian, Canadian	88.5	76.2	131	93	12.28
Theater	80.4	67.6	600	432	12.82

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 Research Area over the last decade in both databases, plus 2021 in *wos\_b\_202204*. Substantial changes between years or differences between the databases may require investigation of the cause.

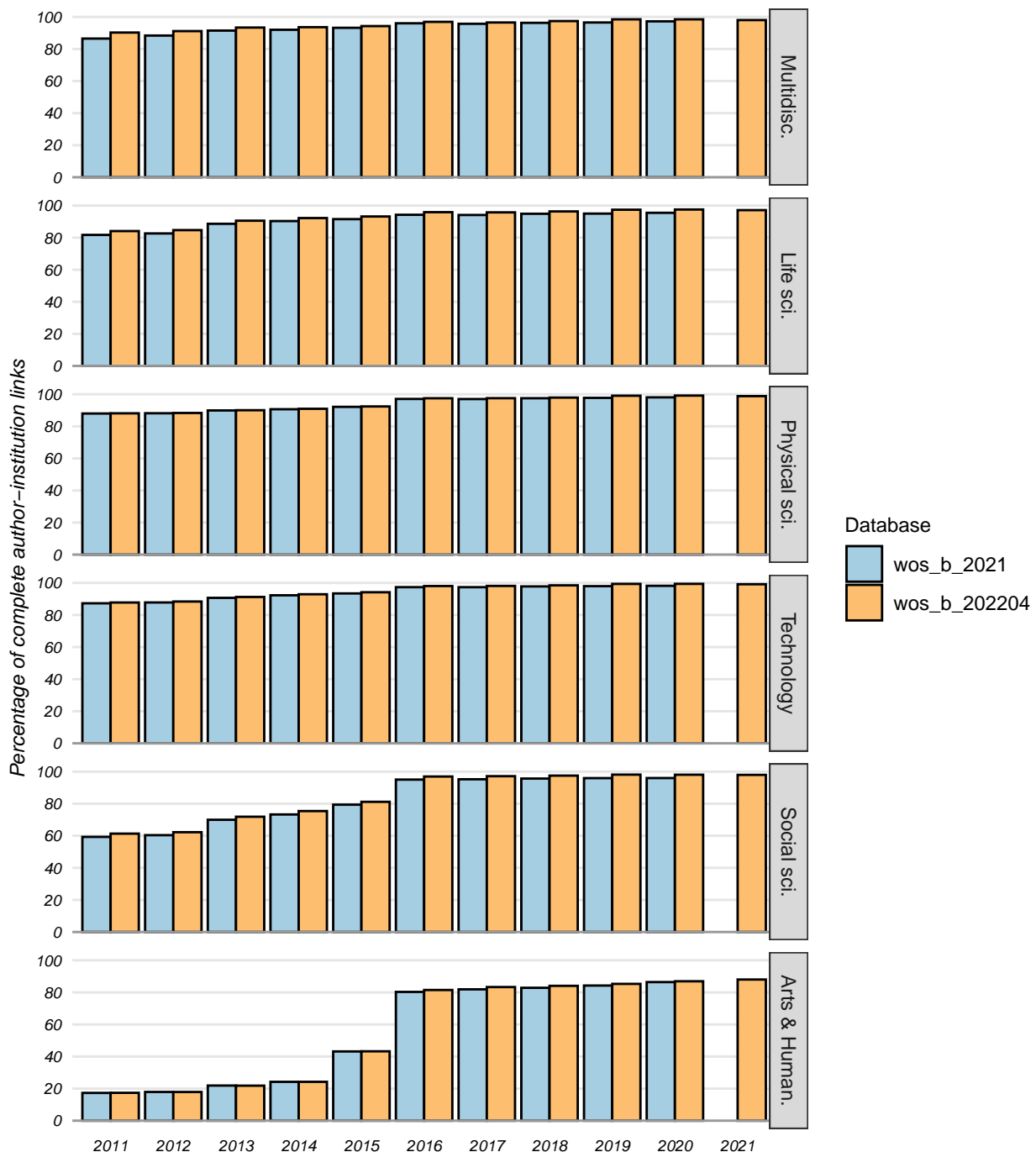


Figure 19: The annual percentage of complete author-institution links by Research Area and database.

## 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 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.

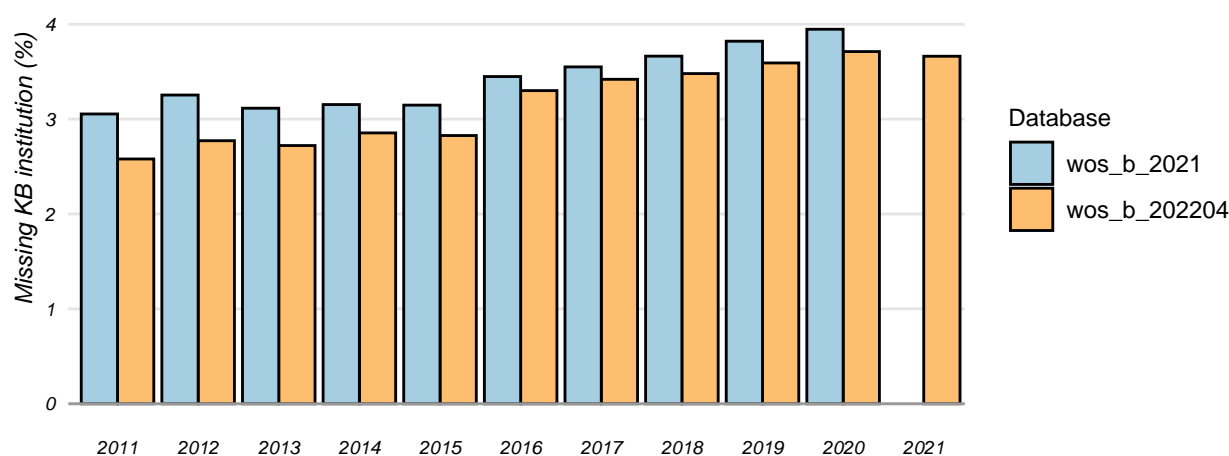


Figure 20: The percentage of German publications in each database that are missing a KB institution.

## 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 6 presents potentially new institutions – these had no publications in 2020 in the wos\_b\_2021 database but more than five publications in 2021 in the wos\_b\_202204 database. Conversely, Table 7 shows the institutions that had at least five publications in 2020 in the wos\_b\_2021 database but no publications recorded in 2021 in the wos\_b\_202204 database. We also highlight in Tables 8 and 9 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 wos\_b\_2021 and wos\_b\_202204 databases.

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

Inst ID	Name	Previous pubs	Current pubs
5617	Max-Planck-Institut für Multidisziplinär	0	430

5616	Leibniz-Institut zur Analyse des Biodive	0	206
5470	Max-Planck-Institut für Verhaltensbiolog	0	174
4700	Helmholtz-Institut Ulm für elektrochemis	0	170
5628	MSH Medical School Hamburg – University	0	164
5610	Helmholtz-Institut Erlangen-Nürnberg für	0	153
4766	Helmholtz-Institut Mainz (HIM)	0	152
5612	Helmholtz-Institut Münster (HI MS)	0	146
4139	Helmholtz-Institut für Pharmazeutische F	0	109
5641	Hearing4all	0	105
5652	MSB Medical School Berlin – Hochschule f	0	81
4765	Helmholtz-Institut Freiberg für Ressourc	0	80
5569	Leipzig Heart Institute GmbH	0	74
5660	Heidelberger Institut für Radioonkologie	0	74
5662	Pettenkofer School of Public Health Münc	0	73
5480	Leibniz-Institut für Resilienzforschung	0	68
5590	IQVIA Commercial GmbH & Co. OHG	0	68
5651	OncoRay – National Center for Radiation	0	65
5634	Einstein Center for Neurosciences	0	60
5614	Helmholtz-Institut Würzburg für RNA-basi	0	56
5611	Helmholtz-Institut für Metabolismus-, Ad	0	54
5595	DBFZ Deutsches Biomasseforschungszentrum	0	48
4764	Helmholtz International Center for FAIR	0	29
5661	MVZ CCB Frankfurt und Main-Taunus GbR	0	29
5615	Fraunhofer Cluster of Excellence Immune-	0	26
5538	Kühne Logistics University – Wissenschaf	0	23
5588	CureVac AG	0	18
5643	vivo international e.V.	0	18
5642	Plansee Composite Materials GmbH	0	15
894	Niedersächsisches Institut für historisc	0	14
5631	JCMwave GmbH	0	13
5636	Studienpraxis Urologie	0	12
5650	MVZ Labor Krone GbR	0	12
5640	UroEvidence	0	11
5657	iOMEDICO AG	0	11
823	Bundeszentrale für gesundheitliche Aufkl	0	10
5472	Leibniz-Institut für Finanzmarktforschun	0	10
5619	Alanus Hochschule für Kunst und Gesellsc	0	9
5620	Sigmund Freud PrivatUniversität Berlin	0	9
5646	DOG – Deutsche Ophthalmologische Gesells	0	8
440	Katharina-Kasper gGmbH Frankfurt/Main	0	7
5572	GeneWerk GmbH	0	7
5639	Pirche AG	0	7
238	Onkologische Schwerpunktpraxis Bielefeld	0	6

5591	Berufsverband der Augenärzte Deutschland	0	6
5613	Helmholtz-Institut für Translationale On	0	6
5635	Leberstiftungs-GmbH Deutschland	0	6

Table 7: Institutions with no publications in 2021 in wos\_b\_202204 that had more than 5 publications in 2020 in wos\_b\_2021.

Inst ID	Name	Previous pubs	Current pubs
1030	Max-Planck-Institut für biophysikalische	245	0
5	Zoologisches Forschungsmuseum Alexander	145	0
1073	Max-Planck-Institut für experimentelle M	111	0
482	Fachkrankenhaus Coswig GmbH - Zentrum fü	6	0
3771	SINTERFACE Technologies	6	0
4694	Zentralinstitut für die kassenärztliche	6	0

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

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
48	Leibniz-Institut für Atmosphärenphysik e	22	53	140.9
4762	Leibniz-Institut für Bildungsverläufe e.	27	55	103.7
1229	Thermo Fisher Scientific Inc.	23	45	95.7
5252	Medizinische Hochschule Brandenburg Theo	126	244	93.7
4758	Max-Planck-Institut für empirische Ästhe	42	81	92.9
547	Hochschule Ravensburg-Weingarten	29	54	86.2
646	Hochschule für angewandte Wissenschaften	21	39	85.7
195	Staatliches Museum für Naturkunde Stuttg	54	96	77.8
652	Hochschule Bochum - University of Applie	35	62	77.1
586	Hochschule Magdeburg-Stendal	52	92	76.9
618	Hochschule für angewandte Wissenschaft u	33	58	75.8
1165	Fraunhofer-Institut für Chemische Techno	32	56	75.0
675	WHU - Otto Beisheim School of Management	75	128	70.7
1022	Max-Planck-Institut für Neurobiologie	36	61	69.4
36	Leibniz-Institut für ökologische Raument	28	47	67.9
452	St.-Antonius-Hospital Eschweiler	33	55	66.7
144	Zeppelin Universität - Hochschule zwisch	34	56	64.7
49	Leibniz-Institut für Agrarentwicklung in	53	87	64.2
134	Hochschule für Angewandte Wissenschaften	117	191	63.2
30	Leibniz-Institut für die Pädagogik der N	53	86	62.3
1476	GBG Forschungs GmbH	28	44	57.1
1637	Zentrum für Rhinologie und Allergologie	34	53	55.9

684	Thüringer Landessternwarte Tautenburg (T	61	95	55.7
486	Klinikum Chemnitz gGmbH	29	45	55.2
826	Deutsche Bundesbank	42	65	54.8
17	Museum für Naturkunde Leibniz-Institut f	176	270	53.4
701	Institut für Solarenergieforschung GmbH	21	32	52.4
572	Hochschule für Technik, Wirtschaft und M	27	41	51.9
666	Hochschule Anhalt - Anhalt University of	41	62	51.2
839	Bundesamt für Strahlenschutz	36	54	50.0
1295	Pfizer Deutschland GmbH	22	33	50.0
33	Leibniz-Institut für Pflanzenbiochemie	83	123	48.2
870	Konrad-Zuse-Zentrum für Informationstech	83	122	47.0
2145	Deutsches Forschungszentrum für Künstlic	77	113	46.8
913	Bayerisches Landesamt für Gesundheit und	71	104	46.5
1202	VOLKSWAGEN AG	46	67	45.7
477	Klinikum Darmstadt	40	58	45.0
56	Fraunhofer-Institut für Optronik, System	38	55	44.7
656	Beuth Hochschule für Technik Berlin	36	52	44.4
5327	Huawei	27	39	44.4
127	Universität Hildesheim	70	100	42.9
1058	Max-Planck-Institut für Infektionsbiolog	61	87	42.6
50	Heinrich-Pette-Institut für Experimentel	72	102	41.7
1048	Max-Planck-Institut für Marine Mikrobiol	92	130	41.3
1642	Airbus GmbH	51	72	41.2
5352	Psychologische Hochschule Berlin	39	55	41.0
1143	Fraunhofer-Institut für Photonische Mikr	22	31	40.9
5120	Deutsches Zentrum für integrative Biodiv	342	479	40.1

Table 9: Institutions with more than 20 publications in 2020 in wos\_b\_2021 that decreased in publication counts by over 40% in 2021 in wos\_b\_202204.

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
643	Fachhochschule Dortmund	37	22	-40.5
743	Deutsches Rotes Kreuz e.V.	139	76	-45.3
192	Zentralinstitut für Seelische Gesundheit	66	32	-51.5
5325	Auditory Valley	83	31	-62.7
4428	Restkategorie Universitäten, Kunst- und	87	22	-74.7

### Authors: Median number of authors by Research 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 sc\_traditional discipline in 2020 in both databases, and in the right panel the median number of authors per discipline in 2020 in the wos\_b\_202204 database compared to 2021 in the wos\_b\_202204 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 10.

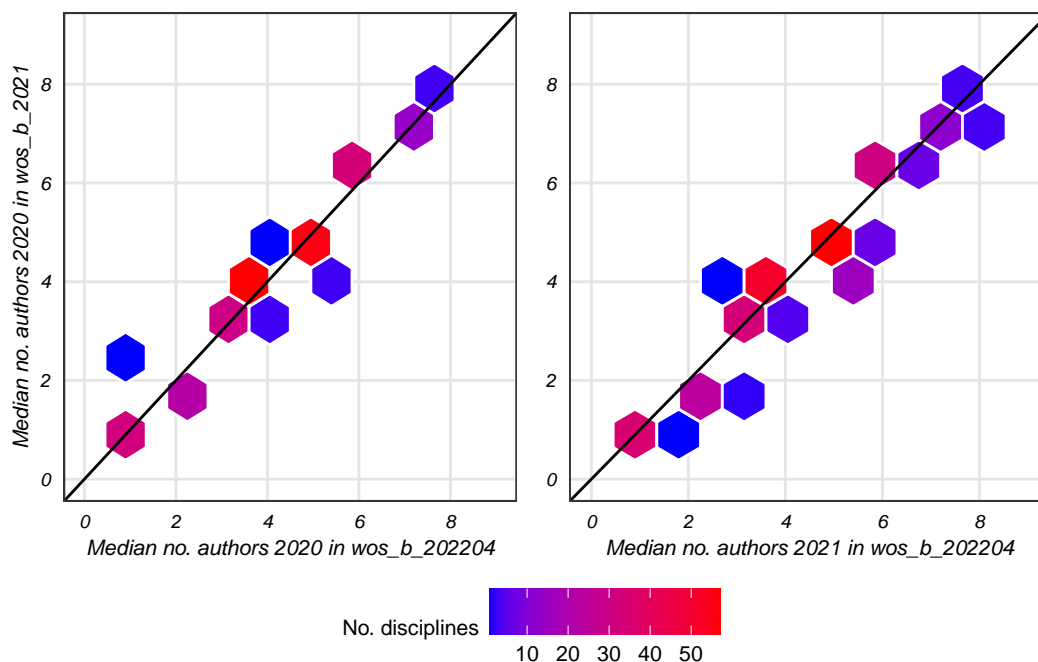


Figure 21: Median number of authors per discipline between databases, where colour denotes the number of disciplines with this combination of median authors.

Table 10: Disciplines where the median number of authors changed by more than 1 between 2020 in wos\_b\_2021 and 2021 in wos\_b\_202204.

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

## Source items: Percentage by Research Area and discipline

Source items refer to whether the publications on the reference list of an indexed publication are also indexed in the database, as opposed to 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 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 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 WoS.

We show in the left-hand panel of Figure 22 the percentage of references that are source items per sc\_traditional 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 wos\_b\_202204 database compared to 2021 in the wos\_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 22, are shown in Table 11. Longer term trends can be seen in Figure 23 where we present the percentage of reference that are source items per Research Area over the last ten common years of both databases.

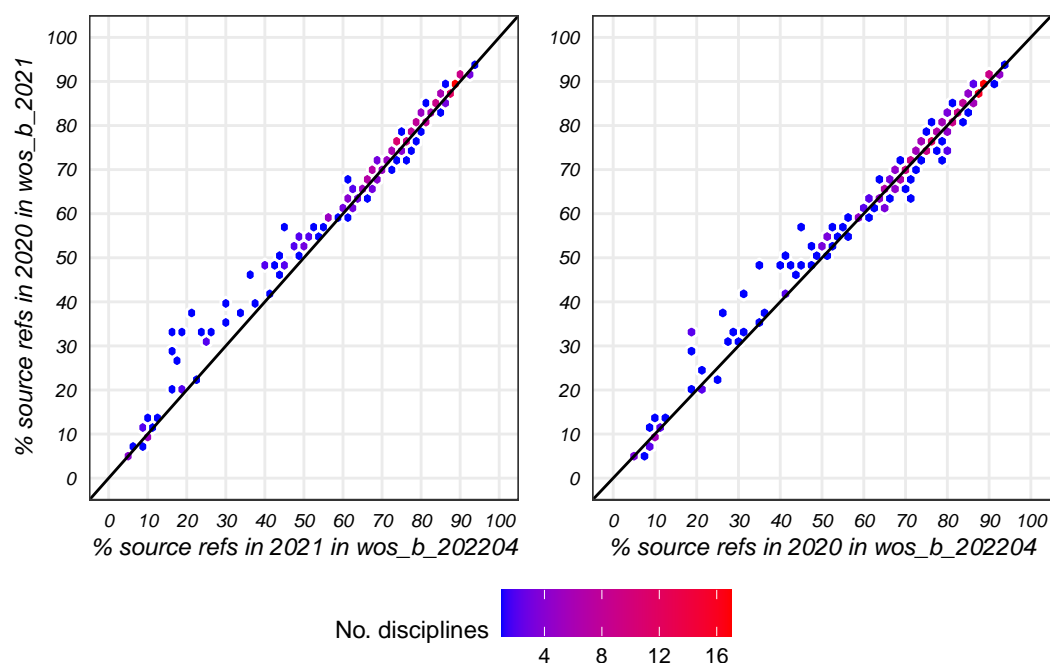


Figure 22: The percentage of cited items that are source items per sc\_traditional discipline by database, where colour denotes the number of disciplines with this combination of source references.

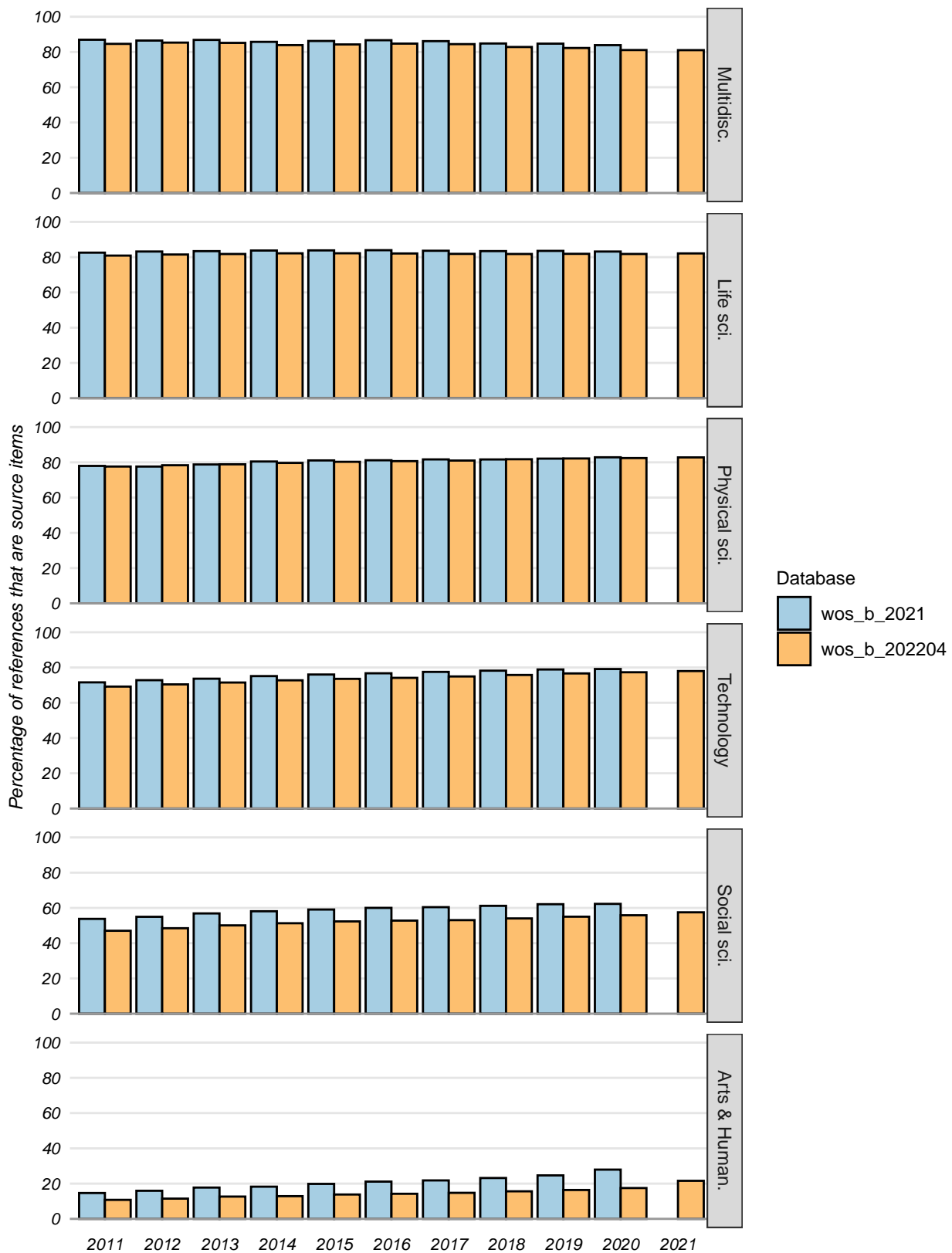


Figure 23: The percentage of references that are source items by Research Area and database over time.

Table 11: Disciplines where the percentage of indexed references changed by 3 or more percentage points between 2020 in wos\_b\_2021 and 2021 in wos\_b\_202204.

Discipline	Prvs % source	Crrnt % source	Change
Mycology	71.7	79.6	8.0
Computer Science, Cybernetics	64.0	71.9	7.9
Computer Science, Hardware & Architecture	62.9	67.4	4.5
Physics, Particles & Fields	75.5	79.6	4.2
Instruments & Instrumentation	74.9	78.8	3.9
Computer Science, Software Engineering	60.4	64.1	3.7
Astronomy & Astrophysics	80.4	83.9	3.5
Robotics	70.0	73.3	3.4
Engineering, Marine	61.8	64.9	3.1
Computer Science, Theory & Methods	61.8	64.8	3.0
Physics, Nuclear	79.1	76.0	-3.0
Urban Studies	54.8	51.7	-3.1
Paleontology	55.8	52.6	-3.2
Philosophy	34.0	30.8	-3.2
Psychology, Experimental	80.4	77.2	-3.2
Demography	53.8	50.4	-3.4
Humanities, Multidisciplinary	25.4	21.9	-3.5
Law	30.6	27.0	-3.6
Social Sciences, Interdisciplinary	54.4	50.6	-3.8
Gerontology	77.0	73.1	-3.9
Social Sciences, Biomedical	67.1	63.0	-4.0
Architecture	33.4	29.1	-4.3
Geography	52.9	48.6	-4.3
Linguistics	49.1	43.5	-5.7
Ethnic Studies	47.9	39.9	-7.9
Anthropology	49.3	41.1	-8.1
Archaeology	40.6	31.5	-9.1
Religion	28.7	18.9	-9.8
Music	38.2	26.7	-11.5
Psychology, Psychoanalysis	47.0	34.9	-12.1
Women's Studies	58.0	45.2	-12.8
Film, Radio, Television	32.0	19.0	-13.0
Art	33.5	19.0	-14.4

## 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.