
Quality assurance at the macro level: Comparing the current and previous WoS snapshots

Dimity Stephen, Stephan Stahlschmidt and Paul Donner

October 2023

Editor:

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

Lange Laube 12 | 30159 Hannover | Germany | info@dzhw.eu | www.dzhw.eu

POB 2920 | 30029 Hannover | Germany

phone: +49 511 450670-0 | fax: +49 511 450670-960

Chairman of the Supervisory Board:

Ministerialdirigent Peter Greisler

Scientific Director:

Prof. Dr. Monika Jungbauer-Gans

Managing Director:

Axel Tscherniak

Registration Court:

Amtsgericht Hannover | HRB 6489

VAT No.: DE291239300

October 2023

Contents

Motivation	1
Set of indicators	1
Set of entities	2
Methodological details	2
Analysis	3
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	27
Source items: Percentage by Research Area and discipline	28

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. 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_202204` and `wos_b_202304` databases is 2021, 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 2019 for the `wos_b_202204` and `wos_b_202304` 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 2021 from the `wos_b_202204` database against 2022 in the `wos_b_202304` 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 (2013)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 a 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 six broad groups: Life Sciences, Physical Sciences, Technology, Social Sciences, Arts and Humanities, and Multidisciplinary. These groups are mapped from the `sc_traditional` disciplines based on a concordance supplied by Clarivate Analytics.

This report is automated. Consequently, blank tables may appear in this report, but they are nonetheless informative about the indicator under examination.

¹Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, 94(3), 851-872. DOI: 10.1007/s11192-012-0775-9

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. Notably, documents may be allocated to multiple types. We have allowed for double-counting between types, in that each document is counted toward each type it is assigned to. However, the total counts each document only once. Figures 2 and 3 show the whole counts of articles and reviews published by selected countries and German sectors over the last 10 years and 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_202304 database contains a greater number of publications for the most recent years than the wos_b_202204 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.

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.

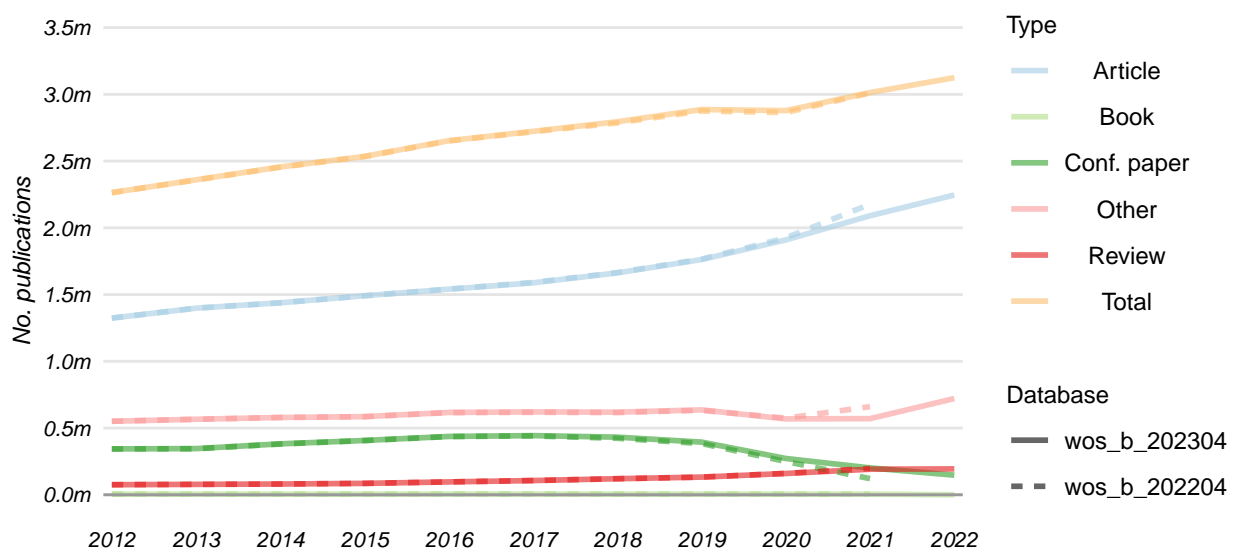


Figure 1: Number of documents in each database over time by type.

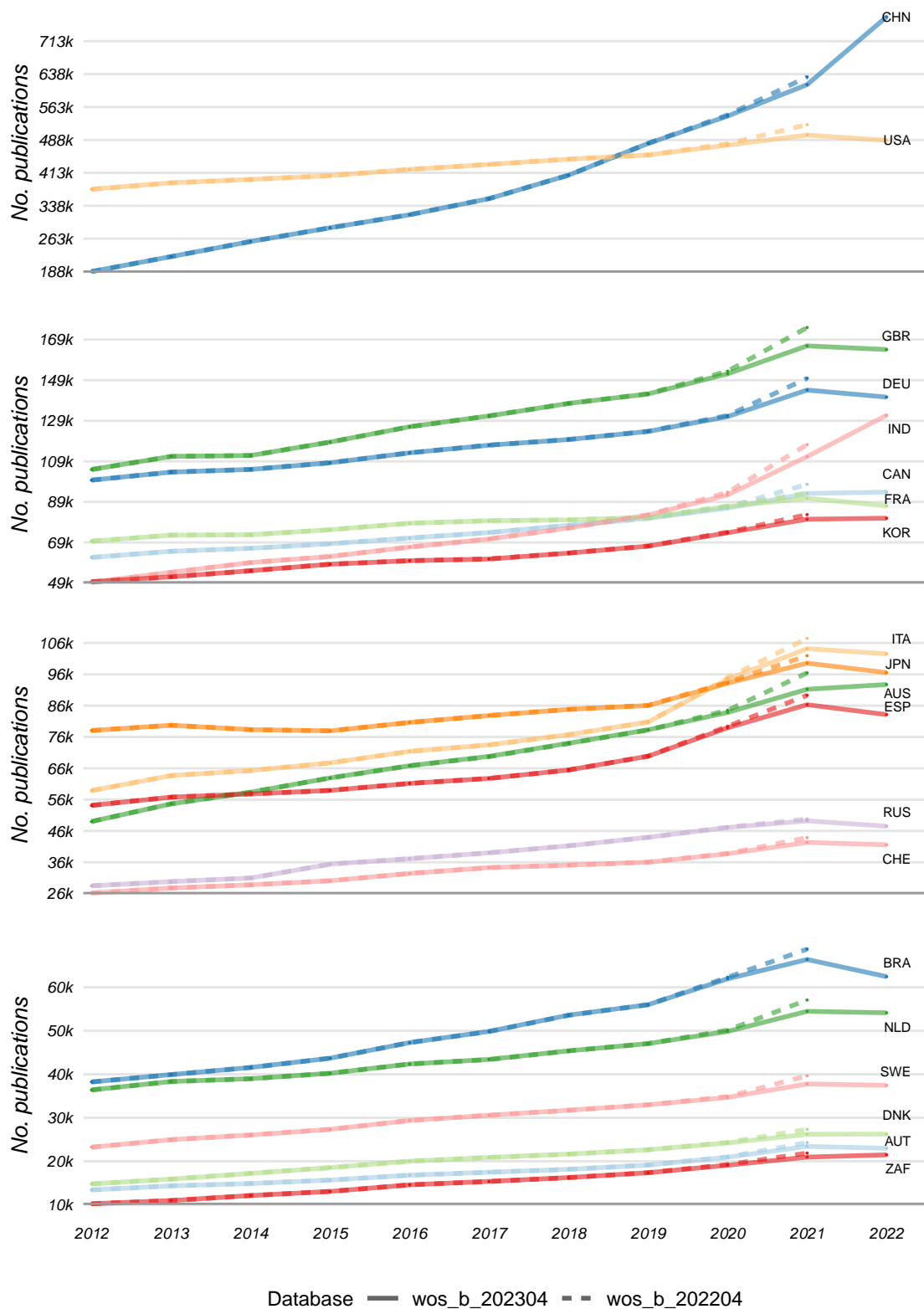


Figure 2: Whole counts of articles and reviews by country and database over time. Please note the panels' different axes.

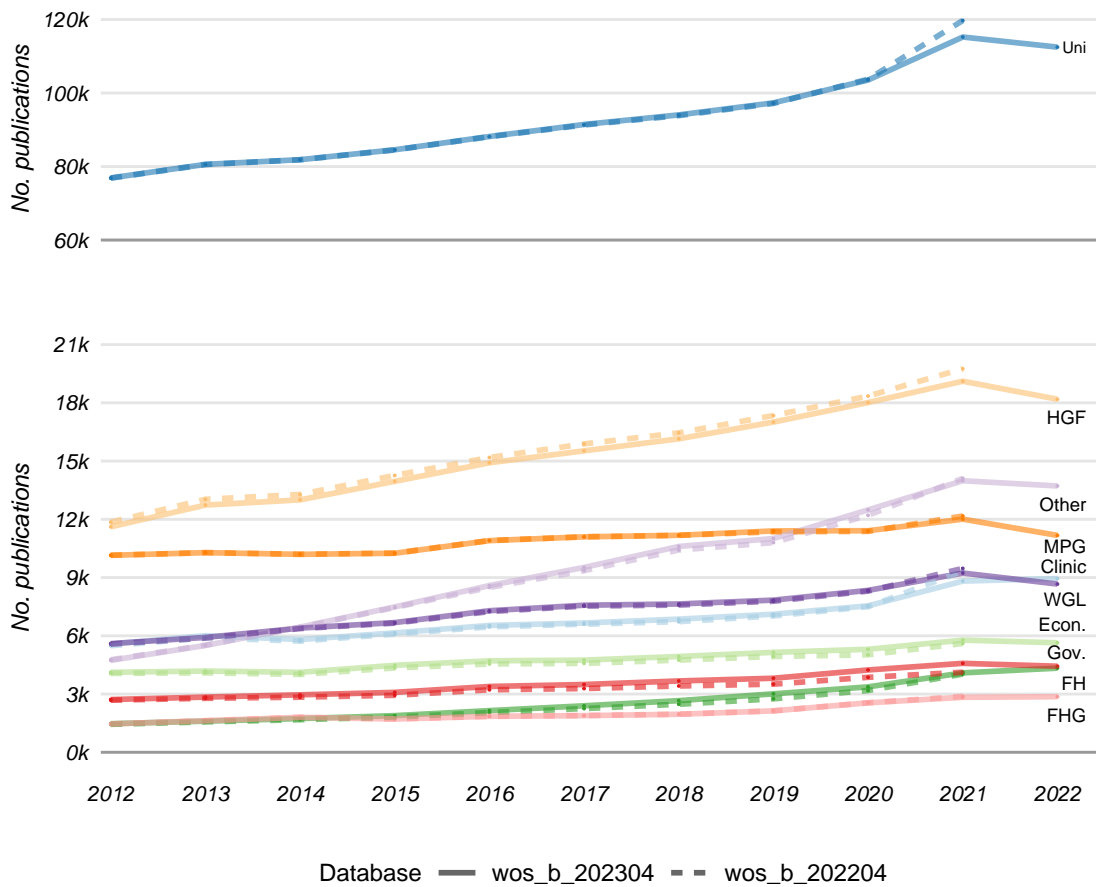


Figure 3: Whole counts of articles and reviews by German sector and database over time. Please note the panels' different axes.

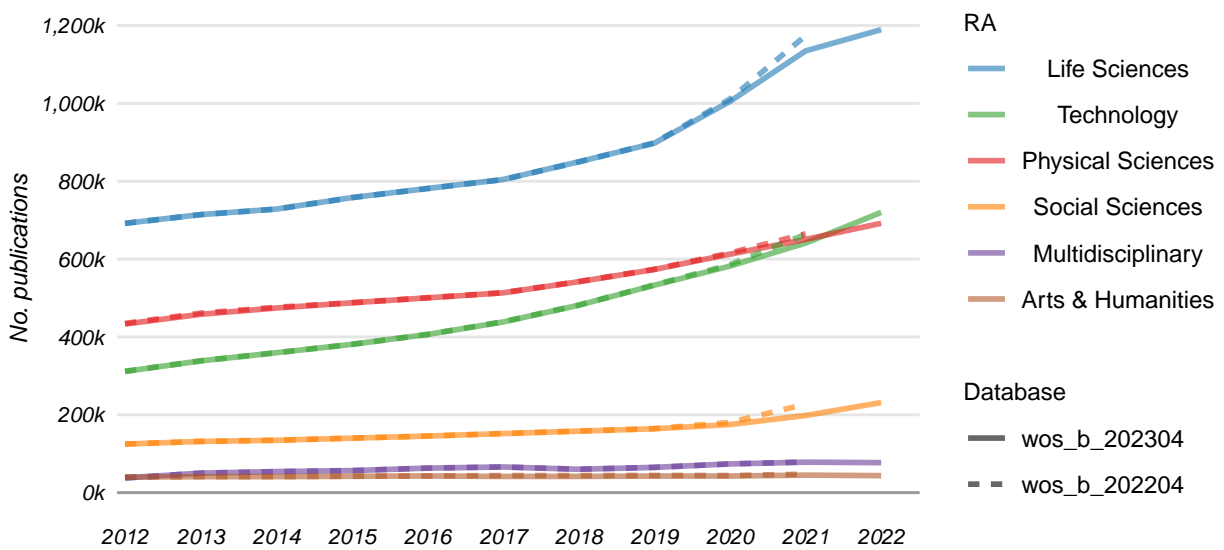


Figure 4: Whole counts of articles and reviews by RA and database over time.

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 2021 in the `wos_b_202204` database to those with 2022 content in the `wos_b_202304` database. Titles were used as all journals have titles recorded, while some journals are missing ISSNs. Titles in `wos_b_202204` but not in `wos_b_202304` were considered removed, while titles in `wos_b_202304` but not in `wos_b_202204` were considered added. In total, 179 journals were added and 227 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.

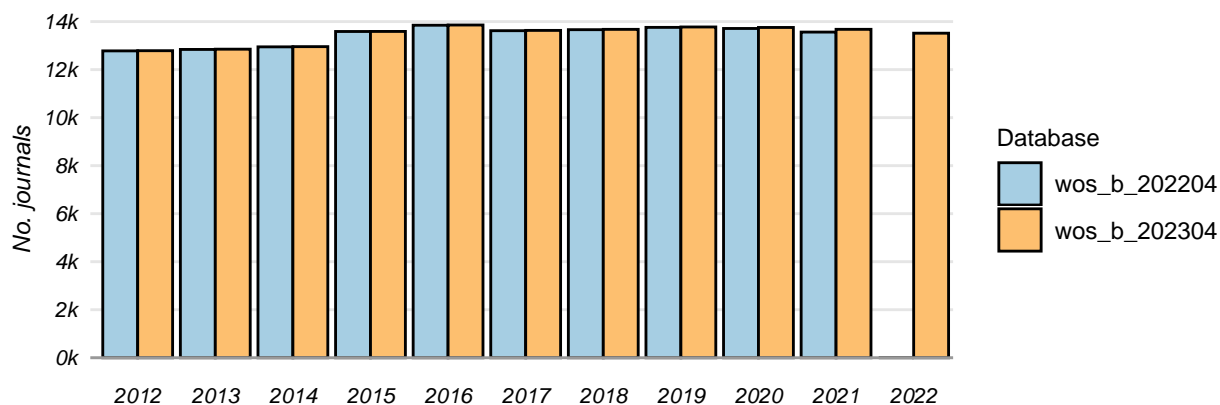


Figure 5: The number of journals indexed in each database over time.

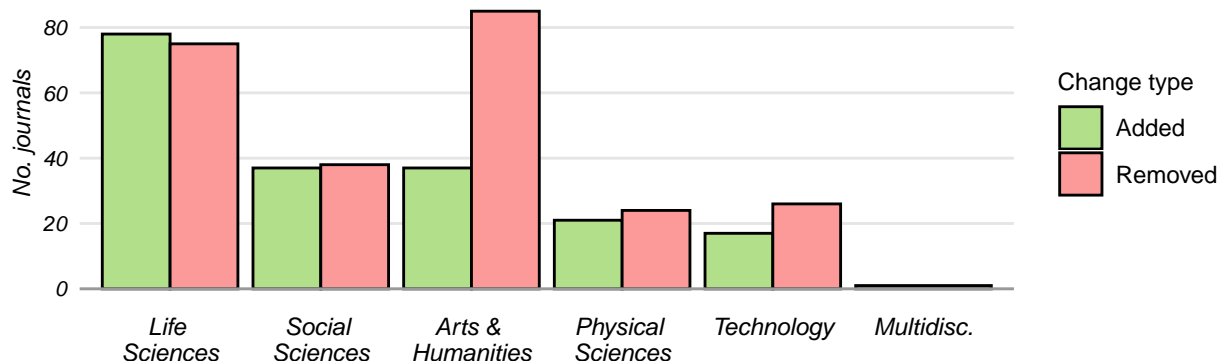


Figure 6: The number of journals added or removed between 2021 in `wos_b_202204` and 2022 in `wos_b_202304` by RA.

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.

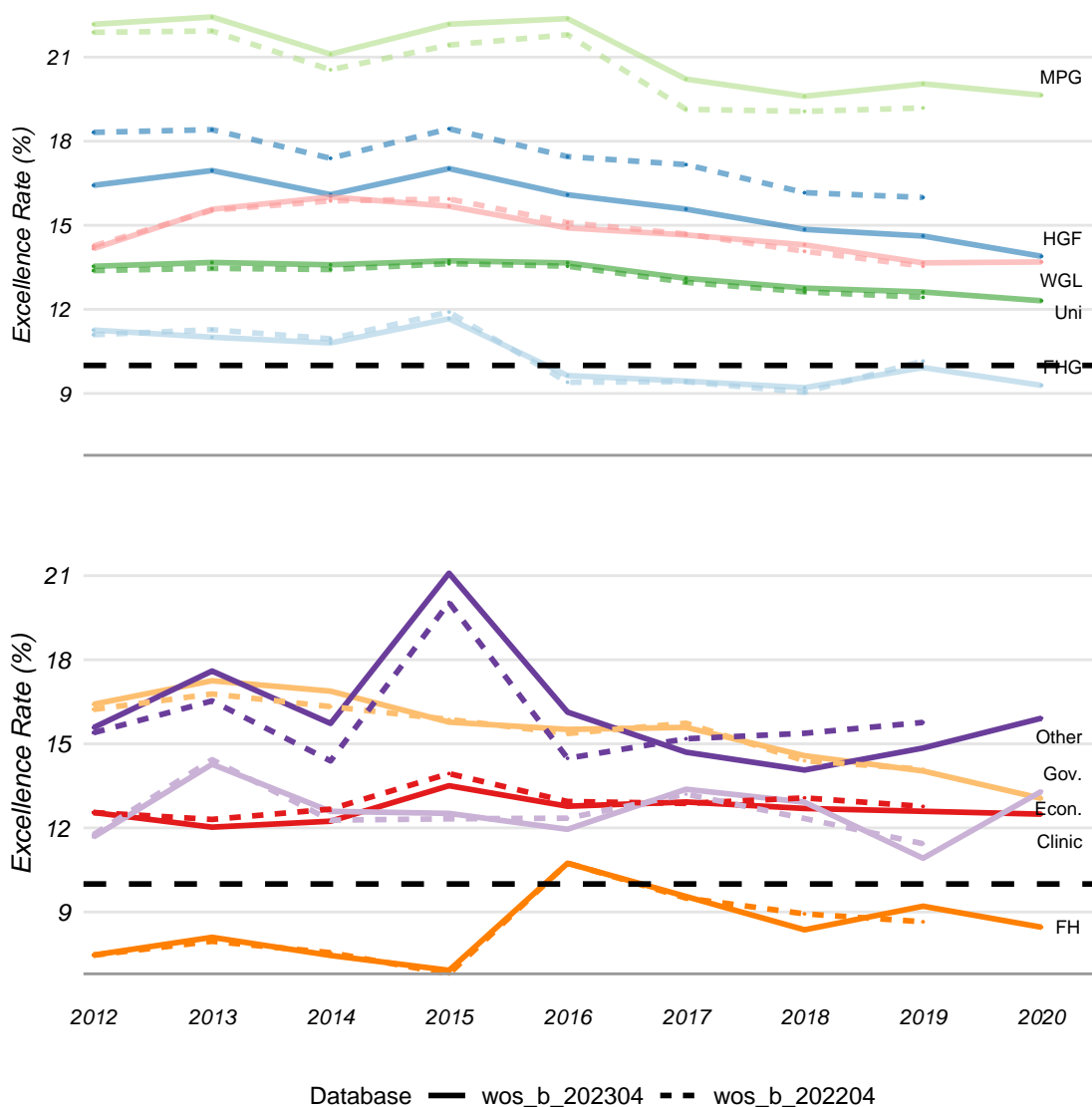


Figure 7: ERs, based on whole counts, by German sector and database over time. The black line is the expected 10% threshold.

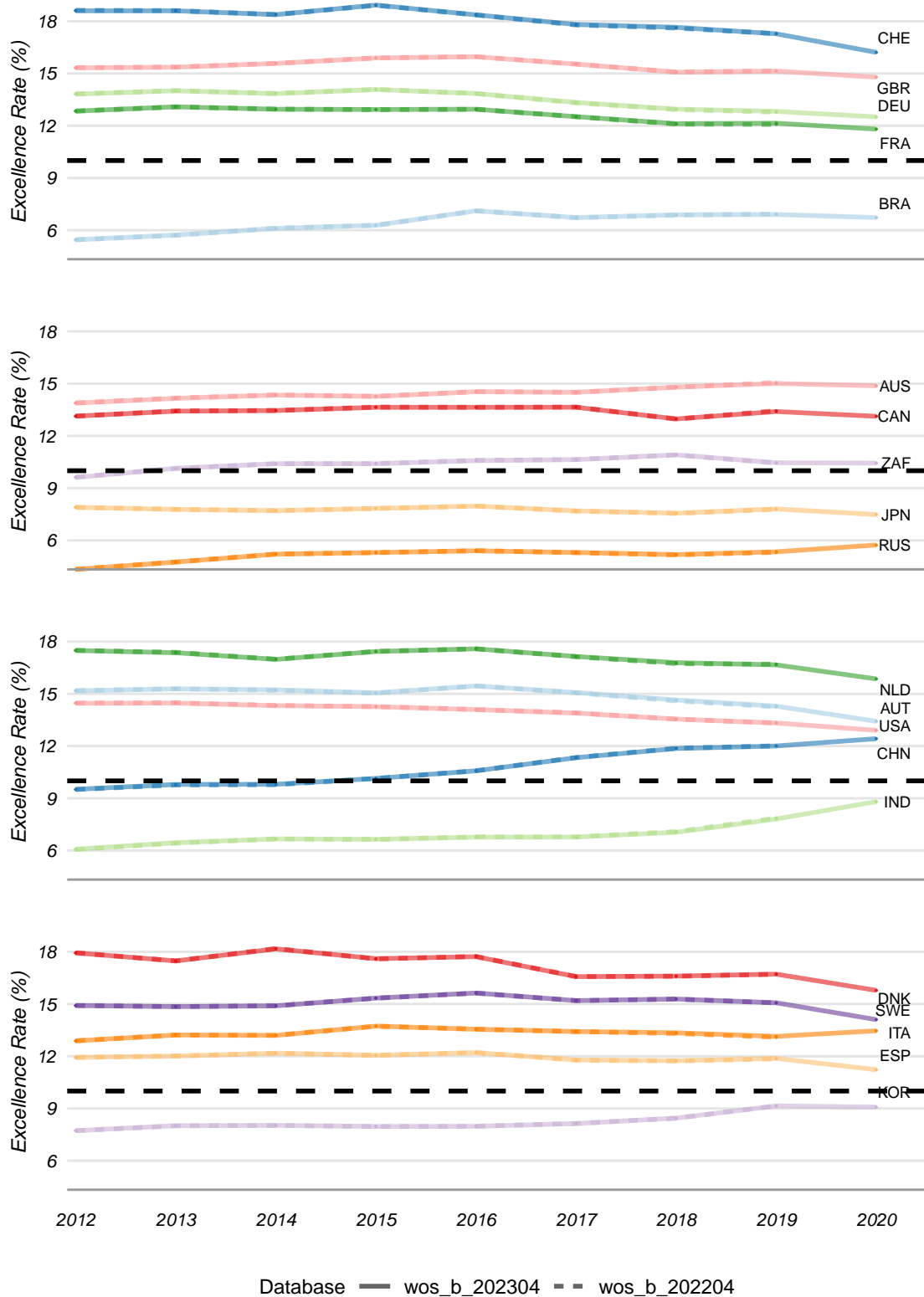


Figure 8: ERs, based on whole counts, by selected country and database over time. The black line is the expected 10% threshold.

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.

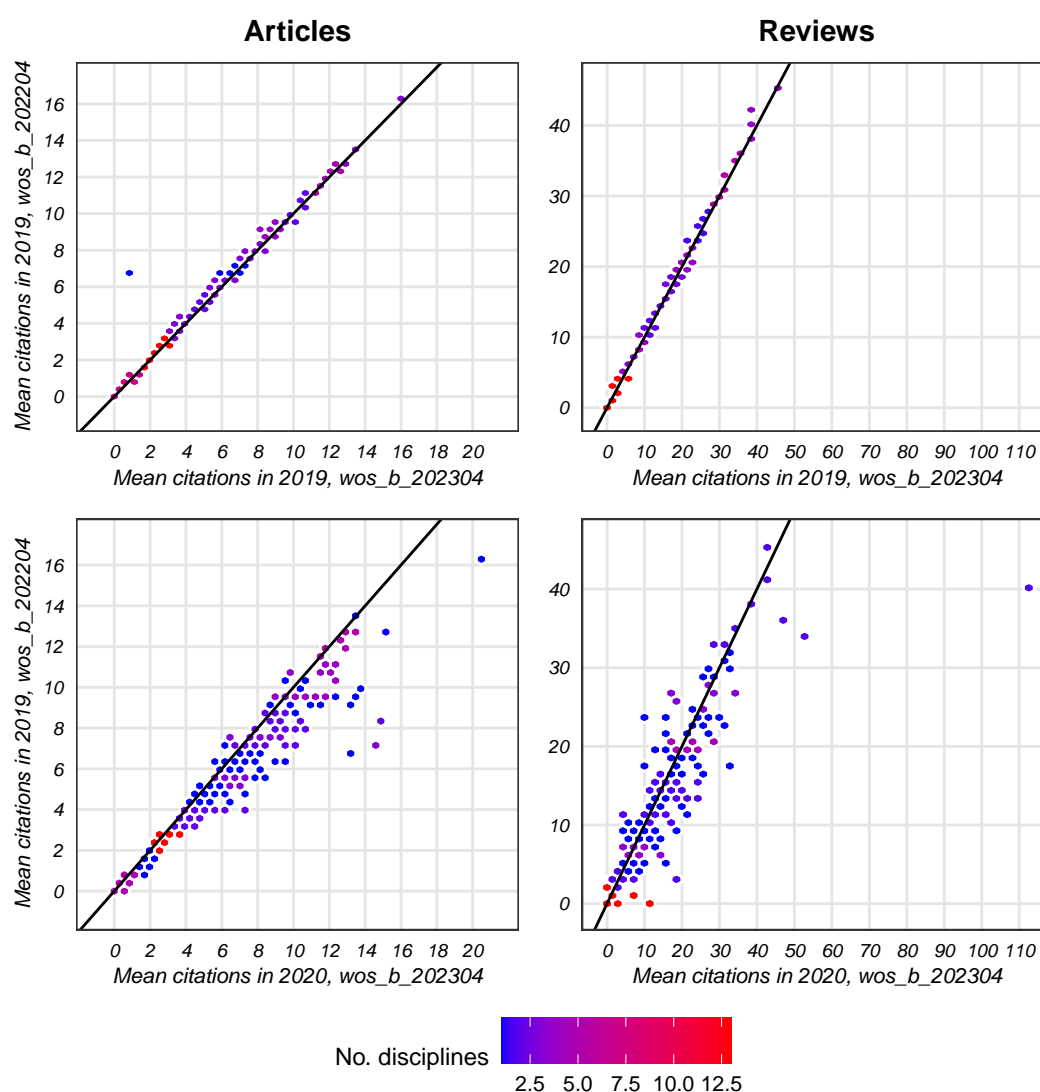


Figure 9: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

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 2019 in wos_b_202204 and 2020 in wos_b_202304, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. currnt pubs.	Perc. diff.
Architecture	0.7	1.5	1372	104.6
Virology	7.3	14.7	3719	101.4
Mining & Mineral Processing	3.9	7.5	271	95.5
Infectious Diseases	6.8	13.0	10040	90.8
Medicine, General & Internal	8.4	14.7	28456	75.5
Medical Laboratory Technology	4.7	7.4	2676	55.8
Public Administration	4.0	6.2	901	54.0
Agricultural Economics & Policy	4.3	6.6	1144	53.5
Public, Environmental & Occupational Health	5.4	8.3	24202	51.8
Computer Science, Hardware & Architecture	6.4	9.4	5931	48.1
Psychology, Biological	6.2	9.1	1743	47.4
Social Sciences, Mathematical Methods	4.0	5.9	379	47.1
Regional & Urban Planning	4.4	6.4	266	45.2
Logic	1.4	2.0	106	43.2
Psychiatry	7.1	10.1	8480	43.1
Immunology	9.4	13.2	16843	41.1
Information Science & Library Science	5.6	7.9	1998	40.8
Business, Finance	5.0	7.0	5936	40.6
Critical Care Medicine	10.0	14.0	4398	40.5
Hematology	9.3	13.1	5815	40.0
Radiology, Nuclear Medicine & Medical Imaging	7.2	9.8	9403	37.0
Industrial Relations & Labor	3.9	5.3	670	36.7
Social Issues	4.1	5.5	1070	36.5
Development Studies	5.4	7.3	2170	35.5
Social Sciences, Interdisciplinary	3.4	4.6	2795	35.4
Mathematics, Interdisciplinary Applications	5.8	7.8	4522	35.2
Psychology, Social	5.8	7.8	2558	34.9
Medical Informatics	5.7	7.7	883	34.4
Materials Science, Characterization & Testing	4.7	6.2	1685	33.8
Hospitality, Leisure, Sport & Tourism	7.9	10.5	3089	33.6
Health Policy & Services	4.0	5.3	2125	32.7
Otorhinolaryngology	4.1	5.4	3893	31.0
Transplantation	6.1	8.0	344	31.0

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 2019 in wos_b_202204 and 2020 in wos_b_202304, or the previous MC3 was 0.

Discipline	Previous MC3	Current MC3	No. crnt pubs.	Perc. diff.
Asian Studies	0.0	0.1	14	Inf
Folklore	0.0	0.3	3	Inf
History Of Social Sciences	0.0	1.5	2	Inf
Literature, American	0.0	0.1	12	Inf
Literature, Slavic	0.0	0.2	11	Inf
Medieval & Renaissance Studies	0.0	0.3	6	Inf
Psychology, Psychoanalysis	0.0	0.4	5	Inf
Social Sciences, Mathematical Methods	0.0	12.5	4	Inf
Imaging Science & Photographic Technology	2.8	17.2	4	516.1
Medical Ethics	1.5	7.0	1	366.7
Materials Science, Characterization & Testing	5.3	15.1	24	186.7
Instruments & Instrumentation	5.1	14.6	110	186.1
Physics, Particles & Fields	39.9	112.0	10	180.8
Linguistics	2.6	6.5	48	152.0
Ergonomics	6.7	13.7	16	105.8
Social Issues	3.7	7.4	27	102.7
Engineering, Marine	9.4	18.9	88	100.9
Humanities, Multidisciplinary	2.1	4.1	74	95.6
International Relations	2.8	5.4	39	91.7
Operations Research & Management Science	11.8	22.1	21	87.8
Nuclear Science & Technology	6.8	12.8	86	87.7
Infectious Diseases	13.2	24.1	1306	82.5
Virology	17.7	32.2	660	82.3
Industrial Relations & Labor	8.4	14.7	20	74.1
Demography	4.4	7.6	10	73.3
Statistics & Probability	2.7	4.6	34	71.7
Communication	5.5	9.3	81	70.1
Materials Science, Textiles	5.2	8.5	55	64.5
Physics, Condensed Matter	15.7	25.4	63	62.1
Women's Studies	4.9	3.8	6	-21.8
Education, Special	7.2	5.6	128	-22.4
Physics, Mathematical	3.9	3.0	47	-22.5
Ornithology	8.2	6.1	19	-25.4
Materials Science, Paper & Wood	25.7	18.4	61	-28.5
Urban Studies	21.8	15.4	8	-29.3
Microscopy	9.3	6.4	18	-31.0
Materials Science, Coatings & Films	20.0	13.4	42	-33.0
Hospitality, Leisure, Sport & Tourism	26.7	17.2	193	-35.7

Physics, Nuclear	23.3	14.7	117	-37.0
Information Science & Library Science	17.8	11.1	115	-37.4
Social Sciences, Interdisciplinary	10.1	6.0	84	-40.4
Archaeology	8.0	4.7	44	-41.4
Public Administration	8.1	4.5	11	-44.7
Engineering, Petroleum	9.0	4.2	4	-52.8
Regional & Urban Planning	23.8	10.0	9	-58.0
Physics, Atomic, Molecular & Chemical	11.1	4.0	25	-64.3
Mineralogy	9.5	3.1	8	-67.1

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

While ERs represent the most highly cited publications and mean citations tell us about what’s average, the percentage of uncited publications can tell us about the entities at the tail end of the citation distribution. When examining uncited publications, we expect to see a decreasing trend in uncited publications over time. This occurs because citation counts are based on the items indexed in each database and so, as 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.

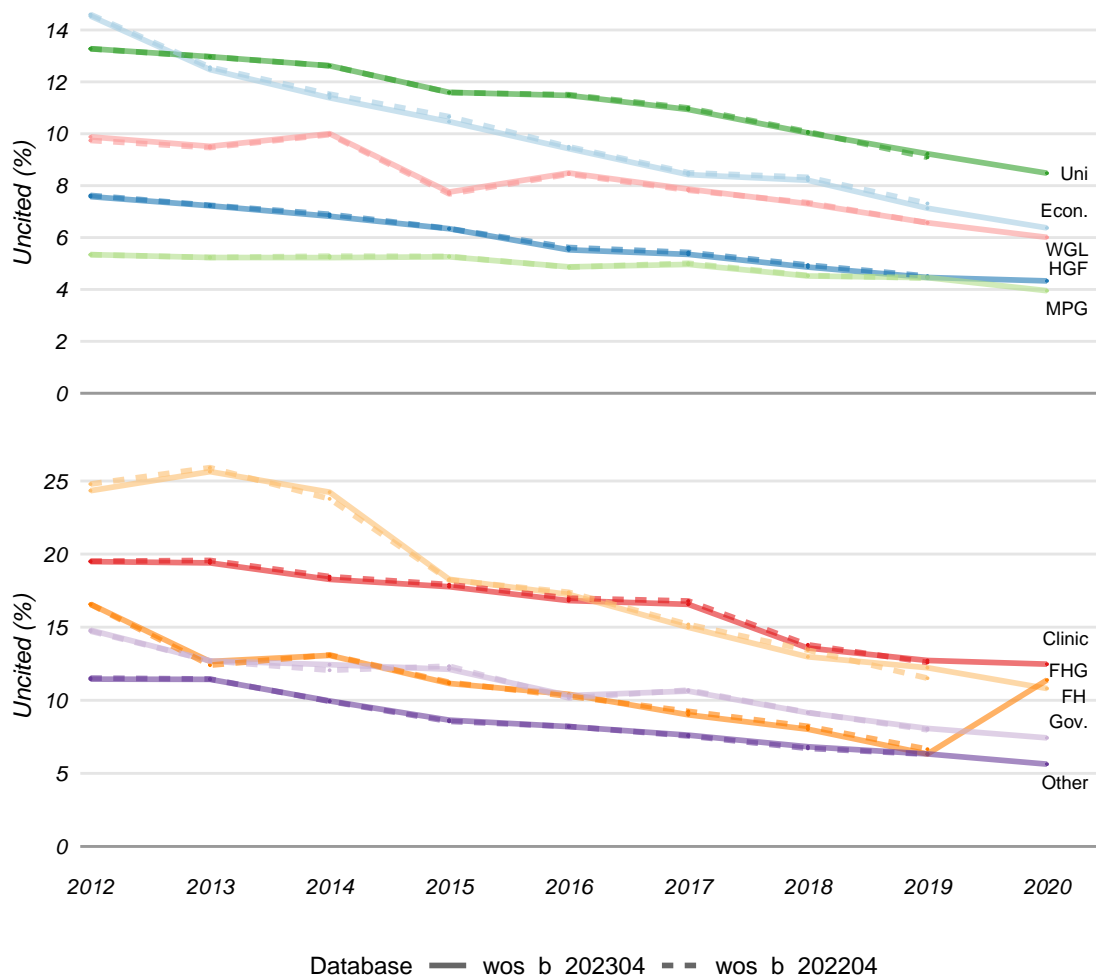


Figure 10: The percentage of uncited publications in each database over time by German sector.

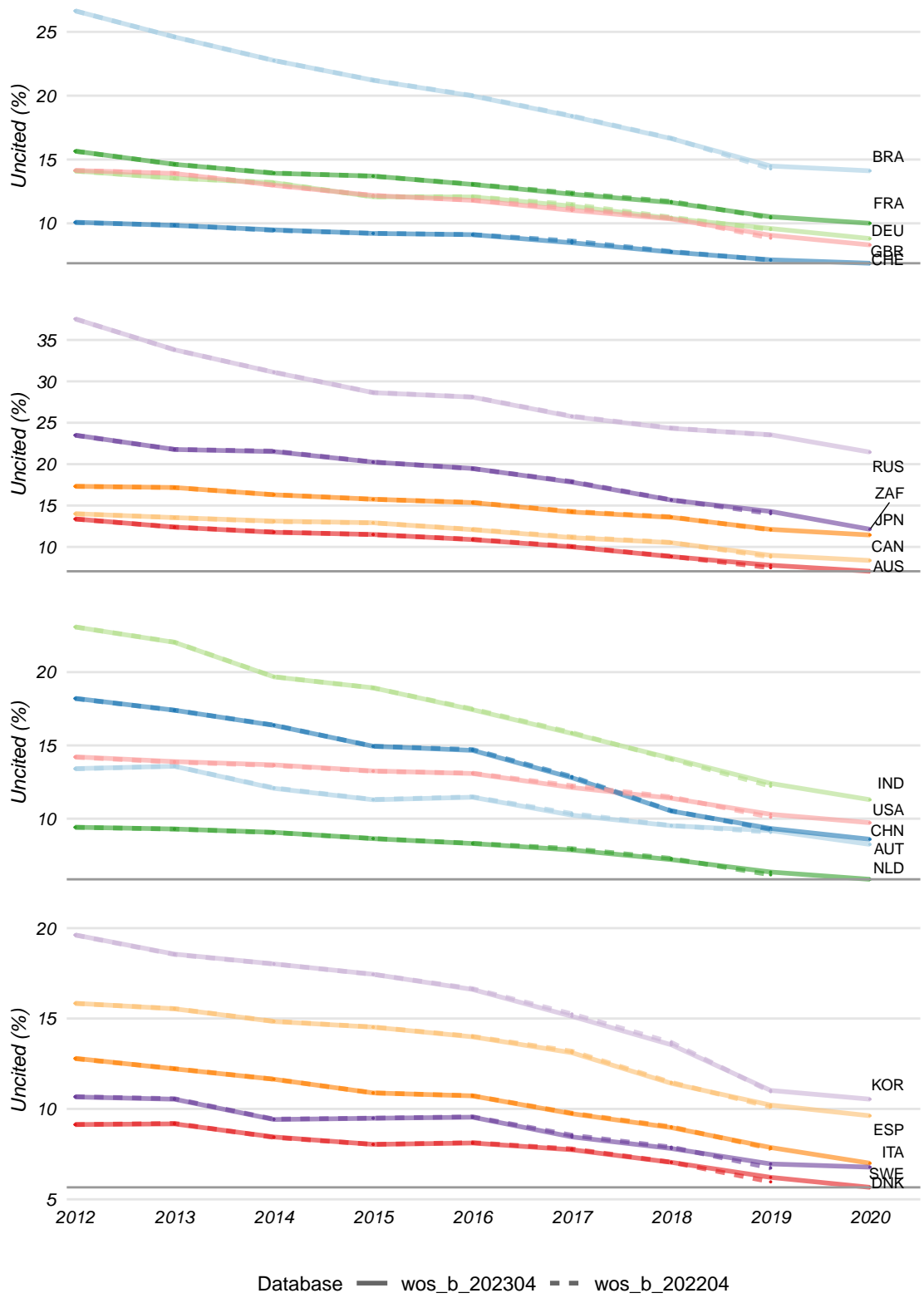


Figure 11: The percentage of uncited publications in each database over time by selected countries.

Disciplines: Changes in discipline classification

This section shows in Table 3 any changes that have been made to WoS' sc_traditional classification. 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.

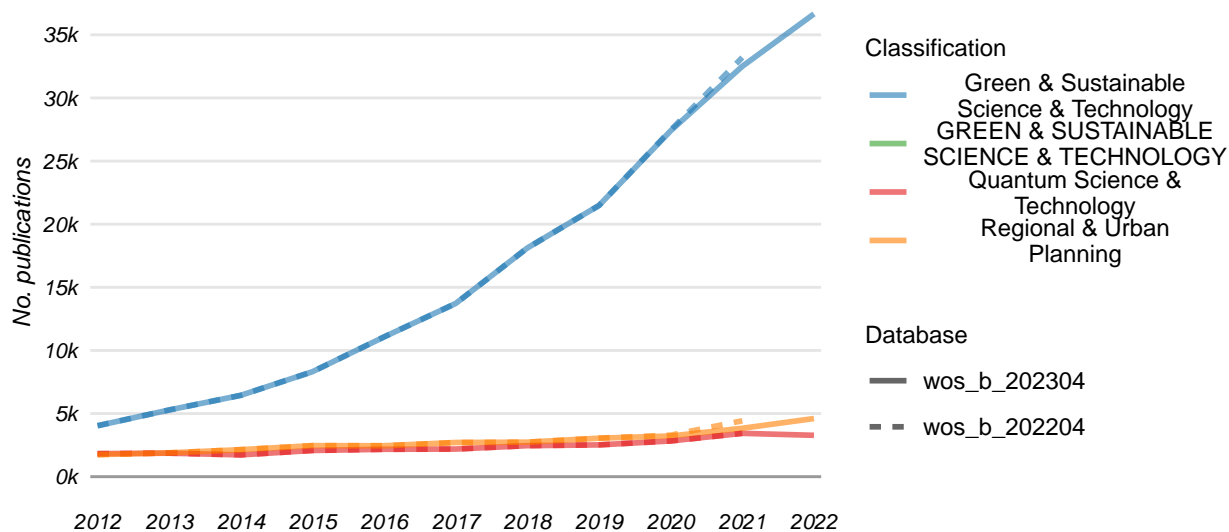


Figure 12: Time-series of sc_traditional disciplines previously observed to have changed.

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 2021 in *wos_b_202204* and 2022 in *wos_b_202304*. The number shown next to each bar is the numerical change in publication counts. We have used whole counting. Disciplines previously identified as being new or removed have not been included here.

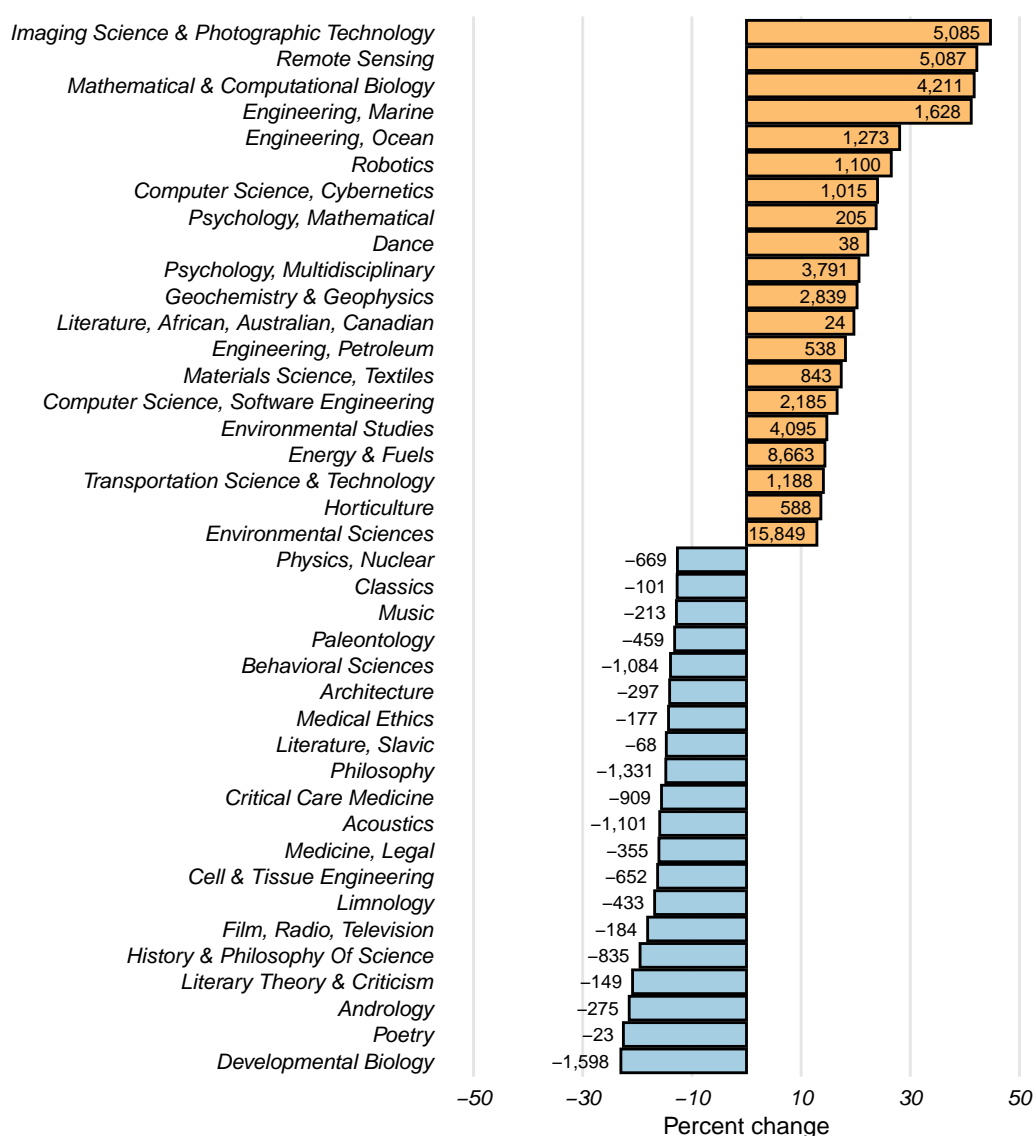


Figure 13: The 40 disciplines with the highest percentage change in publication counts between 2021 in *wos_b_202204* and 2022 in *wos_b_202304*, with numerical difference in counts.

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.

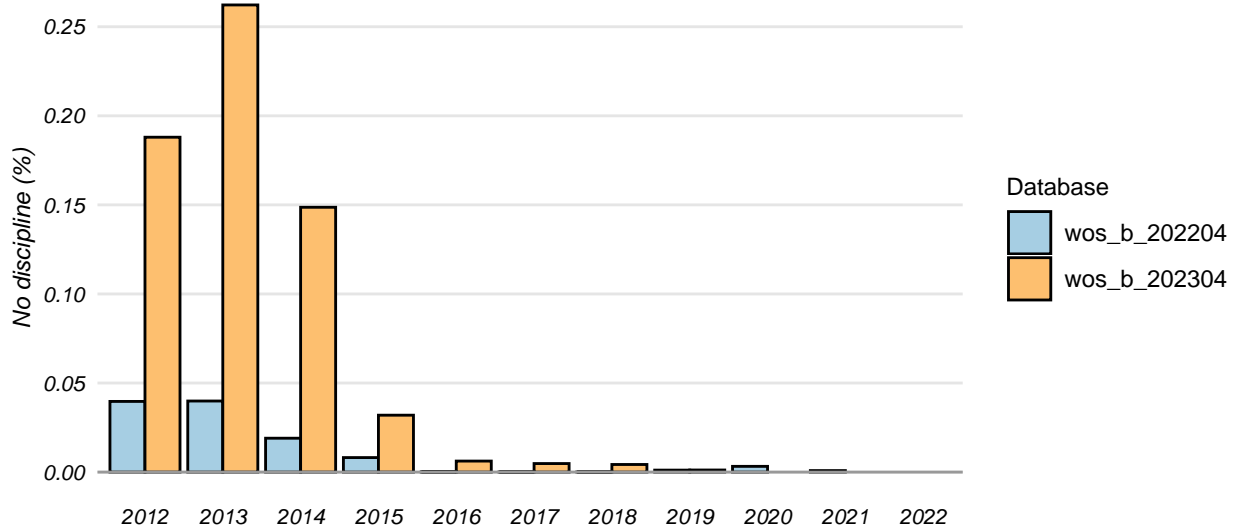


Figure 14: The percentage of publications in each database that do not have a discipline classification.

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_202204 database but not in the wos_b_202304 database (removed items), and items present in the wos_b_202304 database but not in the wos_b_202204 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, 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_202204 and wos_b_202304 databases using the item_id identifier and then calculating the number of items that were added, removed, or had different metadata. The results are shown in Table 4.

```
## Error in eval_tidy(pair$lhs, env = default_env): Objekt 'DOCTYPE' nicht gefunden
```

Table 4: The number of items with changes in metadata between wos_b_202204 and wos_b_202304.

Crrnt year	Prvs year	Diff. year	Diff. pubtype	Diff. doctype	Added	Removed
2016	2016	0	1531	116	0	0
2016	2017	42	0	0	0	0
2016		0	0	0	2140	0
2017	2016	24	0	0	0	0
2017	2017	0	1335	5208	0	0
2017	2021	1	0	1	0	0
2017		0	0	0	3649	0
2018	2017	9	0	0	0	0
2018	2018	0	1405	4721	0	0
2018	2019	8	0	0	0	0
2018	2020	17	0	0	0	0
2018		0	0	0	9578	0
2019	2018	10	0	0	0	0
2019	2019	0	1475	5243	0	0
2019	2020	25	0	3	0	0
2019	2021	1	0	0	0	0
2019		0	0	0	12637	0
2020	2019	132	0	28	0	0
2020	2020	0	0	5198	0	0
2020	2021	185	0	25	0	0
2020		0	0	0	32386	0
2021	2019	49	0	48	0	0
2021	2020	867	0	412	0	0
2021	2021	0	0	7601	0	0
2021		0	0	0	120642	0
2022	2019	689	0	688	0	0
2022	2020	13413	0	13348	0	0
2022	2021	92905	0	92023	0	0
2022		0	0	0	3018216	0
	2016	0	0	0	0	461
	2017	0	0	0	0	985
	2018	0	0	0	0	470
	2019	0	0	0	0	634
	2020	0	0	0	0	3243
	2021	0	0	0	0	25851

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, ISSHP), 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_202204` and `wos_b_202304` databases between 2012 and 2022.

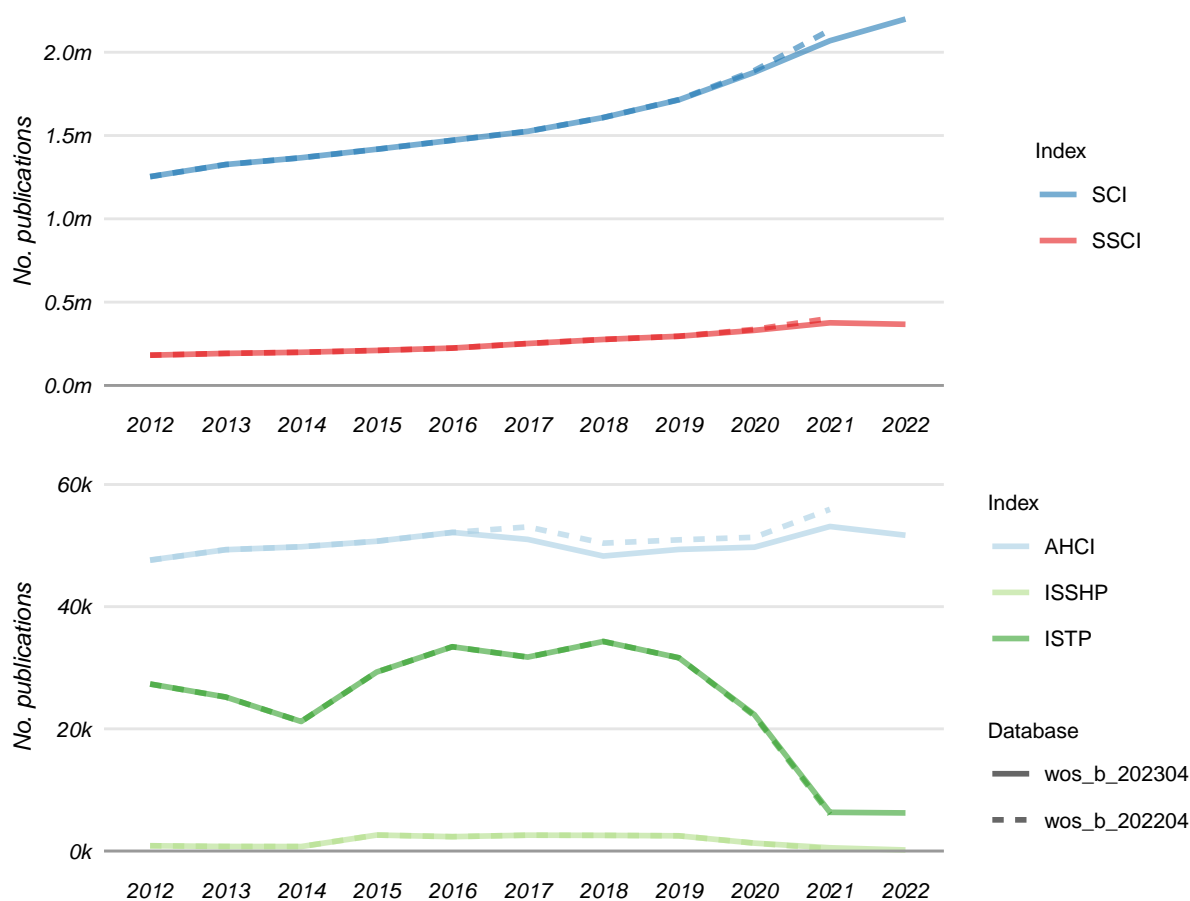


Figure 15: The number of articles and reviews in each WoS index by database over time.

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.

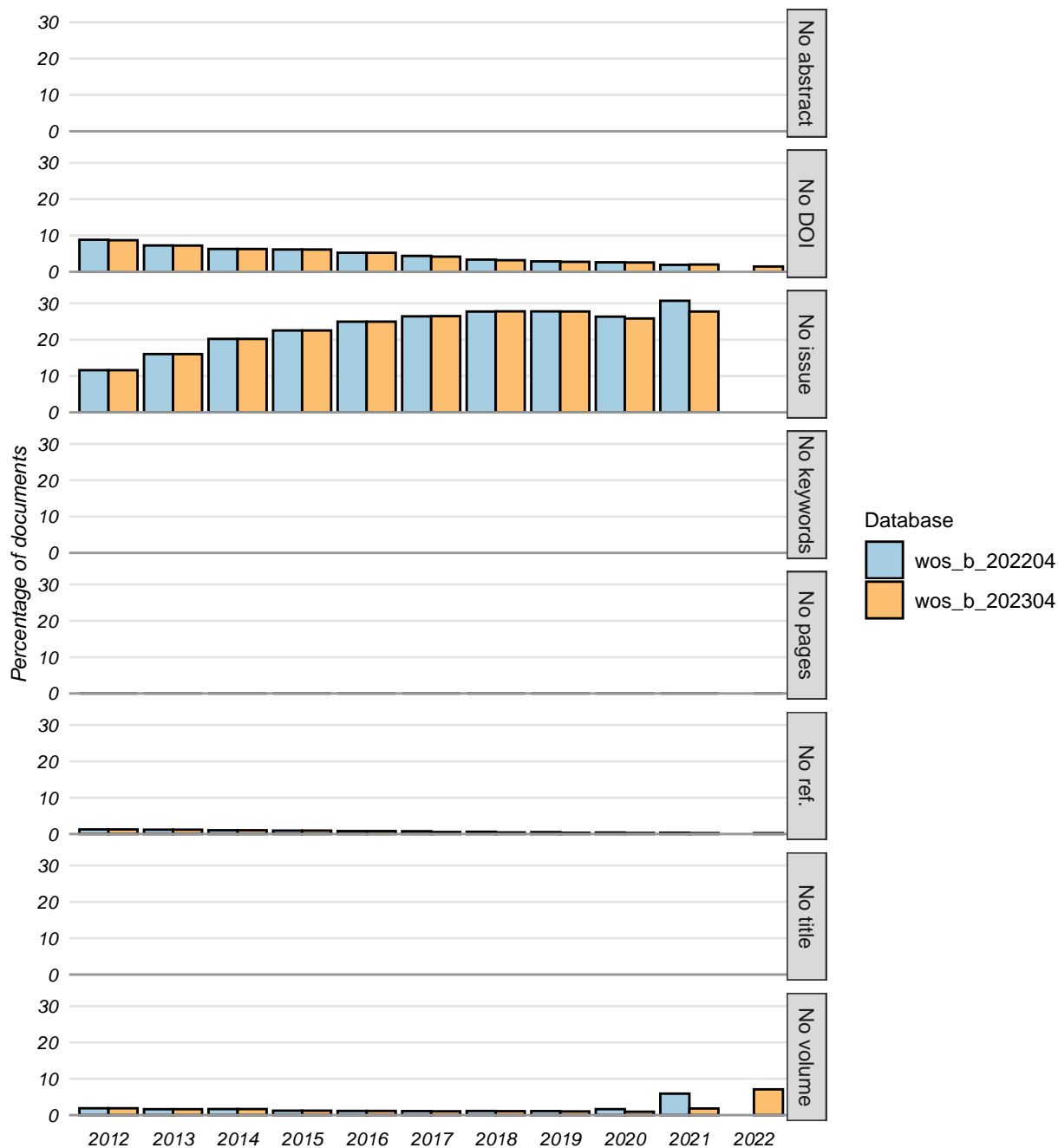


Figure 16: The percentage of items with missing metadata over time by database.

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 items_affiliations tables. We have operationalised missing institution information here as publications that appear in the items table but have no corresponding information in the items_affiliation tables. We present in the top panel of Figure 17 the number of items in each database between 2012 and 2022 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.

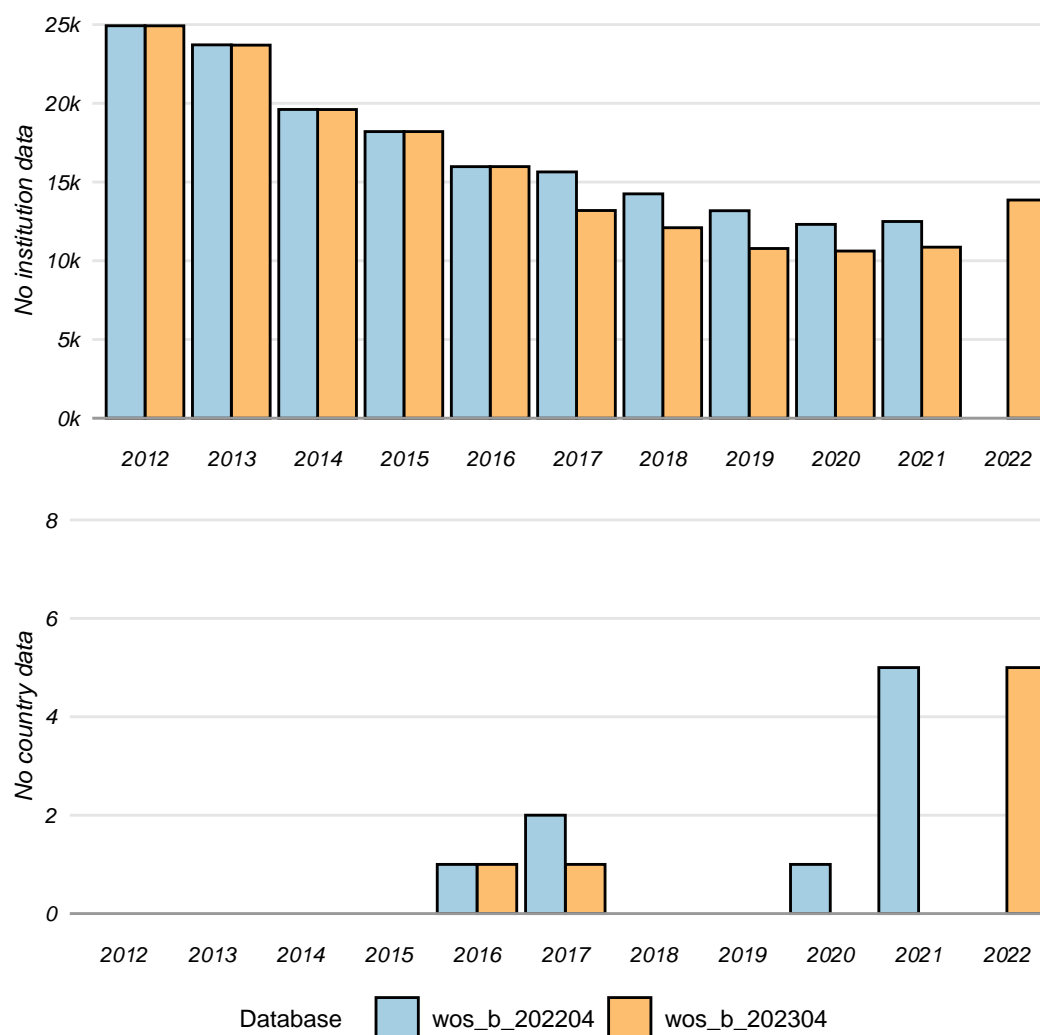


Figure 17: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

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 2021 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 2021 in the wos_b_202204, now compared with the 2022 in the wos_b_202304, indicating potential changes between the latest year in each database.

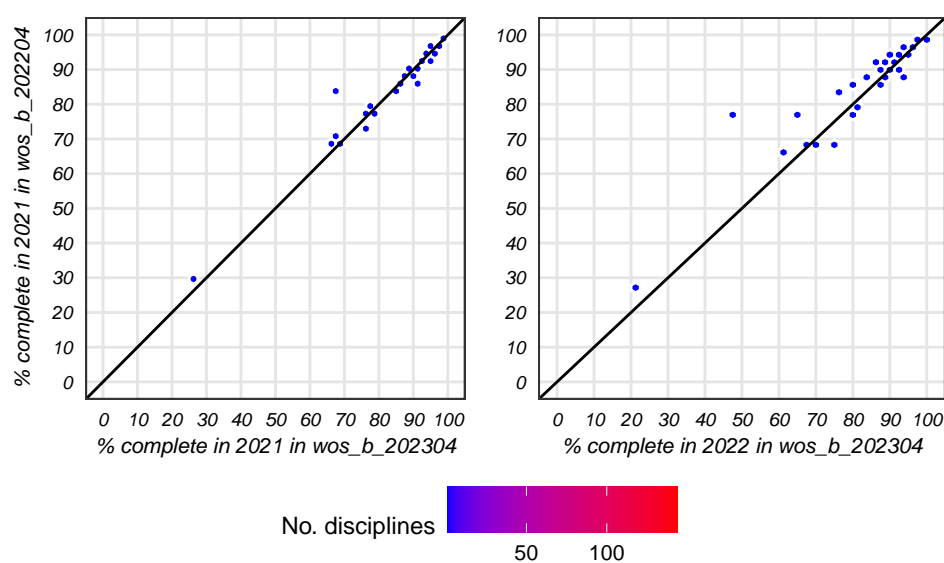


Figure 18: The percentage of complete author-institution links by disciplines.

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 2021 in wos_b_202204 and 2022 in wos_b_202304.

Discipline	Prvs %	Crrnt %	Prvs no.	Crrnt no.	Change
Architecture	68.1	75.6	1436	1370	-7.52
Literary Reviews	27.3	20.0	343	231	7.29
Literature, British Isles	84.2	75.3	267	247	8.93
Literature, African, Australian, Canadian	76.2	65.1	93	95	11.16
Poetry	76.5	48.1	78	38	28.37

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 2022 in wos_b_202304. Substantial changes between years or differences between the databases may require investigation of the cause.

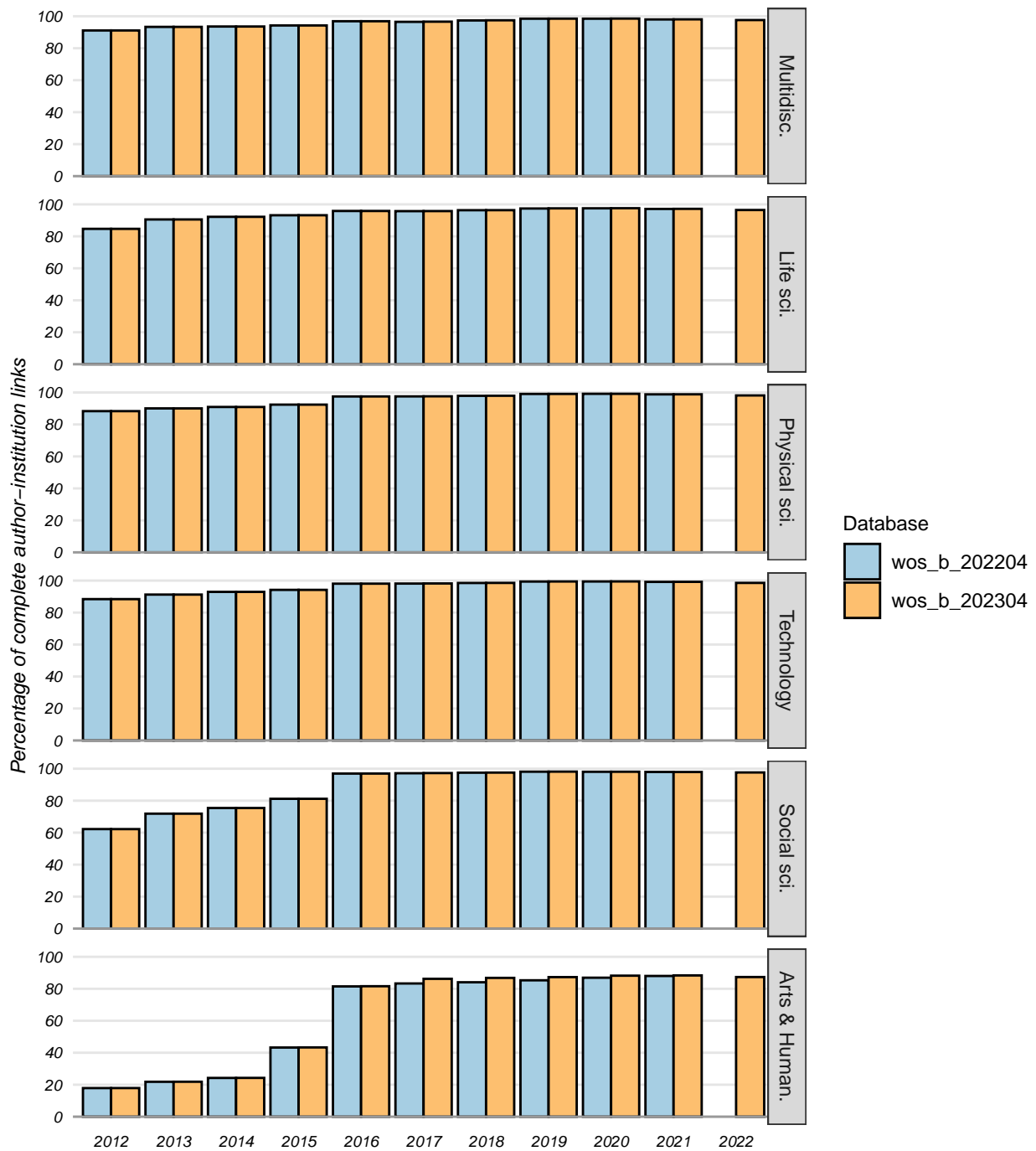


Figure 19: The annual percentage of complete author-institution links by Research Area and database.

German institutions: German publications missing from KB institution coding

In Figure 20 we show the annual percentage of German publications, i.e. those where the German indicator is TRUE, that were not assigned a KB institution code through the institution coding 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.

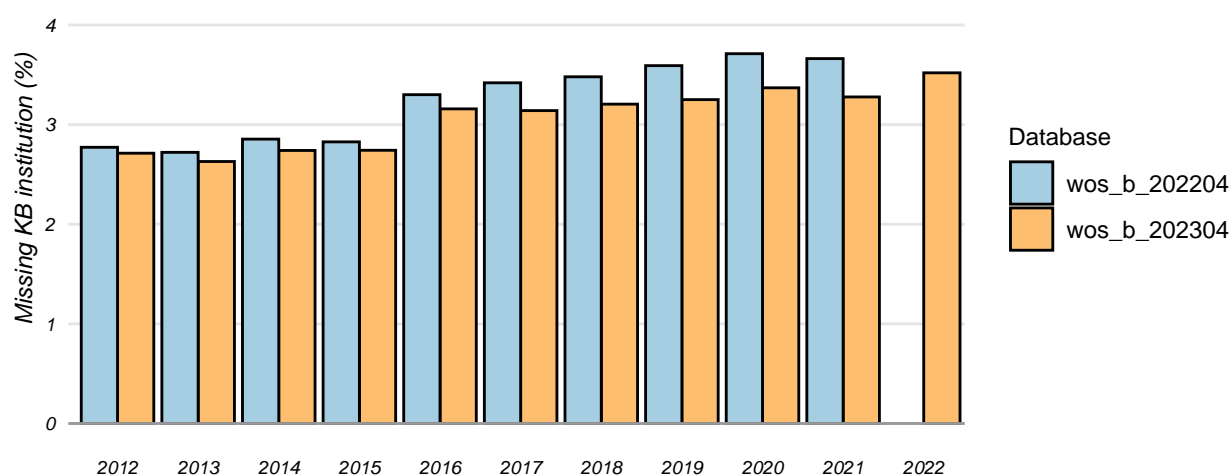


Figure 20: The percentage of German publications in each database that are missing a KB institution.

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 2021 in the wos_b_202204 database but more than five publications in 2022 in the wos_b_202304 database. Conversely, Table 7 shows the institutions that had at least five publications in 2021 in the wos_b_202204 database but no publications recorded in 2022 in the wos_b_202304 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 2021 and 2022 in the wos_b_202204 and wos_b_202304 databases.

Table 6: Institutions with more than 5 publications in 2022 in wos_b_202304 that had no publications in 2021 in wos_b_202204.

Inst ID	Name	Previous pubs	Current pubs
5721	Kerckhoff-Klinik GmbH	0	89
5741	Comprehensive Cancer Center Erlangen-EMN	0	87

5674	BMW GROUP	0	82
979	Hochschule Geisenheim	0	81
5728	LOEWE-Zentrum für Translationale Biodive	0	77
5679	German Institute of Development and Sust	0	59
5714	Hochschule Ruhr West	0	46
5772	Leibniz-Institut für Immuntherapie	0	45
3442	Cochrane Deutschland Stiftung (CDS)	0	42
5715	Hochschule für Gesundheit	0	30
5668	Max-Planck-Zentrum für Physik und Medizi	0	29
5747	Zentrum für internationale Bildungsvergl	0	29
5696	Einstein Center Digital Future	0	27
5732	InfectoGnostics Forschungscampus Jena e.	0	27
5738	Labor Berlin – Charité Vivantes GmbH	0	23
5684	nutriCARD - Der Kompetenzcluster für Ern	0	20
5733	Weizenbaum-Institut e. V.	0	18
5671	Hochschule Hamm-Lippstadt	0	16
5745	Saurierwelt Paläontologisches Museum	0	16
5200	Restkategorie Privatadressen	0	15
5736	BioNTech SE	0	15
5727	LifeGlimmer GmbH	0	14
5746	Techniker Krankenkasse	0	14
5777	Covestro AG	0	14
5695	Lipotype GmbH	0	13
5705	EUFH – Hochschule für Gesundheit, Sozial	0	11
5740	Life Molecular Imaging GmbH	0	11
5725	Global Energy Interconnection Research I	0	10
5750	SRH Hochschule für Gesundheit Gera	0	10
5731	Barkhausen Institut gGmbH	0	9
5753	Hochschule für Wirtschaft und Umwelt Nür	0	9
5759	Alexander von Humboldt Institut für Inte	0	9
5744	Endokrinologiepraxis am Stuttgarter Plat	0	8
5680	Intel Deutschland GmbH	0	7
5726	AIO-Studien-gGmbH	0	7
5739	MRI.TOOLS GmbH	0	7
5749	SCHWIND eye-tech-solutions GmbH	0	7
5770	Max-Planck-Forschungsstelle für Neurogen	0	7
5469	Max-Planck-Institut für Sicherheit und P	0	6
5658	Infektiologikum	0	6
5681	HaaPACS GmbH	0	6

Table 7: Institutions with no publications in 2022 in wos_b_202304 that had more than 5 publications in 2021 in wos_b_202204.

Inst ID	Name	Previous pubs	Current pubs
1583	Bayerische Motoren Werke Aktiengesellsch	83	0

Table 8: Institutions with more than 20 publications in 2021 in wos_b_202204 that increased in publication counts by over 40% in 2022 in wos_b_202304.

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
1541	Daimler AG	21	70	233.3
1022	Max-Planck-Institut für Neurobiologie	61	141	131.1
693	Laser Zentrum Hannover e.V. (LZH)	46	102	121.7
5325	Auditory Valley	31	67	116.1
1047	Max-Planck-Institut für Mathematik	151	283	87.4
5432	Centogene AG	26	47	80.8
5550	Institute for Advanced Sustainability St	26	43	65.4
1041	Max-Planck-Institut für Mikrostrukturphy	85	139	63.5
1971	Technische Hochschule Köln	75	115	53.3
103	Universität der Bundeswehr München	194	291	50.0

Table 9: Institutions with more than 20 publications in 2021 in wos_b_202204 that decreased in publication counts by over 40% in 2022 in wos_b_202304.

Inst ID	Name	Previous pubs	Current pubs	Perc. diff.
---------	------	---------------	--------------	-------------

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 2021 in both databases, and in the right panel the median number of authors per discipline in 2021 in the wos_b_202204 database compared to 2022 in the wos_b_202304 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.

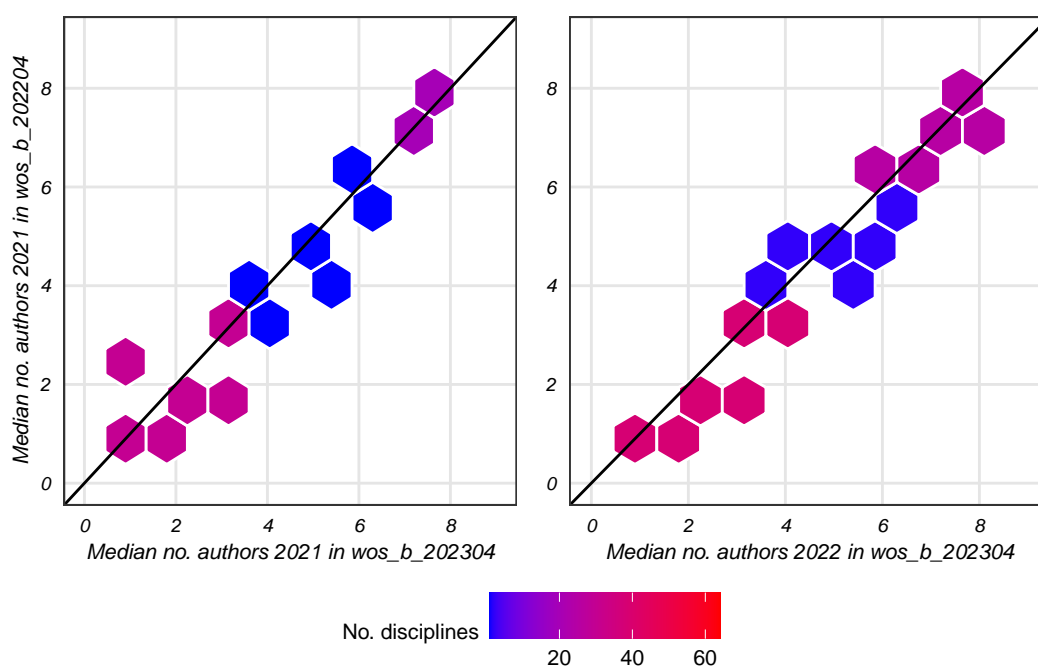


Figure 21: Median number of authors per discipline between databases, where colour denotes the number of disciplines with this combination of median authors.

Table 10: Disciplines where the median number of authors changed by more than 1 between 2021 in wos_b_202204 and 2022 in wos_b_202304.

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 2021 in both databases, and in the right-hand panel the percentage of references that are source items per discipline in 2021 in the *wos_b_202204* database compared to 2022 in the *wos_b_202304* 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.

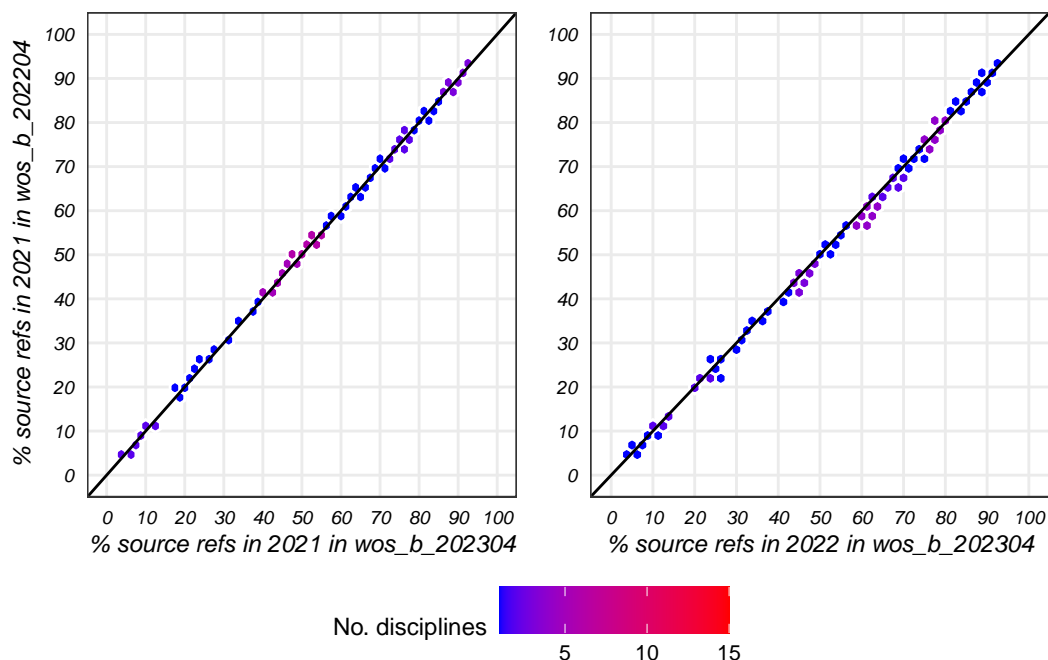


Figure 22: The percentage of cited items that are source items per *sc_traditional* discipline by database, where colour denotes the number of disciplines with this combination of source references.

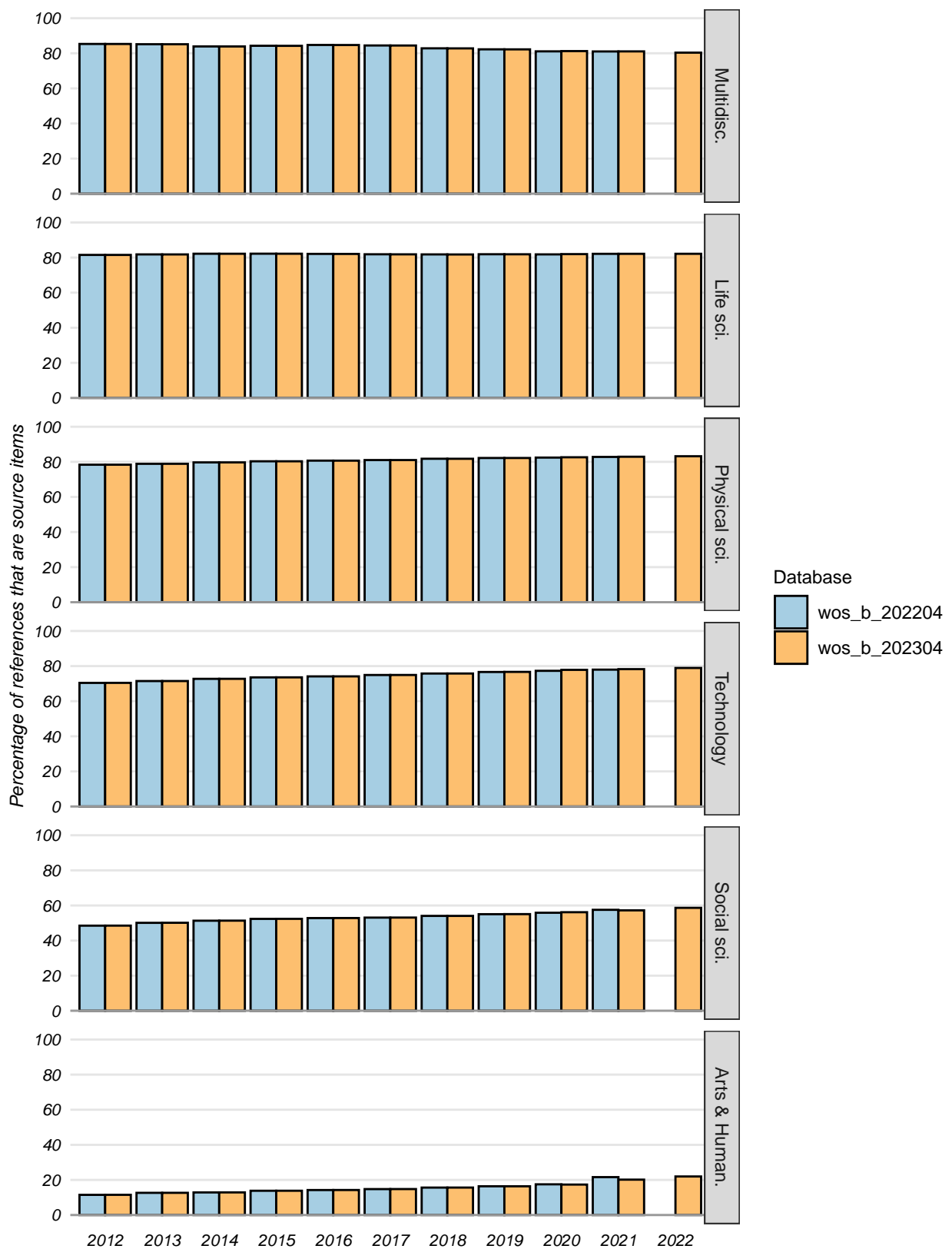


Figure 23: The percentage of references that are source items by Research Area and database over time.

Table 11: Disciplines where the percentage of indexed references changed by 3 or more percentage points between 2021 in wos_b_202204 and 2022 in wos_b_202304.

Discipline	Prvs % source	Crrnt % source	Change
Agricultural Economics & Policy	56.5	61.5	5.0
Engineering, Petroleum	59.0	63.5	4.5
Engineering, Marine	64.9	68.4	3.5
Humanities, Multidisciplinary	21.9	25.4	3.5
Dance	21.1	24.3	3.2
Regional & Urban Planning	55.6	58.8	3.2
Law	27.0	22.7	-4.2