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

Dimity Stephen, Stephan Stahlschmidt and Paul Donner

Report on wos_b_2020 and wos_b_2021

September 2021

DZHW
German Centre for Higher Education Research and Science Studies

Editor:
German Centre for Higher Education Research and Science Studies (DZHW) GmbH
Lange Laube 12 | 30159 Hannover | Germany | info@dzhw.eu |www.dzhw.eu
POB 2920 | 30029 Hannover | Germany
phone: +49 511 450670-0 | fax: +49 511 450670-960
Chairman of the Supervisory Board:
Ministerialdirigent Peter Greisler
Scientific Director:
Prof. Dr. Monika Jungbauer-Gans
Managing Director:
Karen Schlüter
Registration Court:
Amtsgericht Hannover | HRB 6489
VAT No.: DE291239300
September 2021

Contents

Motivation

The aim of the report is to identify any potential changes in data between or within database versions that may indicate quality issues. To do so it offers:

- a visual comparison
- between time-series over the last 10 years
- stemming from the current and previous KB database snapshots
- on several key indicators
- for national, sectoral and institutional entities.

The DZHW already conducts quality assurance testing at the micro-level for KB bibliometric databases before the tables enter the production environment. This testing is invaluable to ensuring tables and variables contain the expected content. This report supplements the current micro-level approach by examining changes in key variables between the latest two iterations of the databases at the macro-level of institutions, sectors, countries, disciplines.

This report is not an exhaustive analysis of the databases' content, nor does it investigate any anomalies identified within the databases. However, this report probes the core variables fundamental to common bibliometric analyses, serves as an overview of the current state of the databases, and highlights changes that may indicate issues with data quality that warrant further investigation to understand or rectify. Changes may arise through several means. For instance, the database provider may add or remove journals from indices, change the discipline classification, or change how the classification is applied. The KB may identify new or decommissioned institutions, which can affect publication output for particular disciplines, or countries may implement policies regarding publication practices that can exert a substantial influence on the content published over time. This report aims to provide users of the KB databases with an overview of potential changes soon after the databases enter the production environment, so that these factors may be considered in analyses.

Set of indicators

The indicators we have chosen reflect the core variables in the database that are fundamental to key bibliometric analyses and indicators. We provide context to the selection of variables and what information can be determined from their examination in each of the following sections.

We make two sets of comparisons in this report. For indicators where it is important to consider trends over time, such as whole publication counts, we compare the databases for the 10 years up to the year for which both have complete data. For example, the latest common year with complete data for the wos_b_2020 and wos_b_2021 databases is 2019, as data for the absolute latest year in each database are incomplete. Similarly, where citation-based indicators are used, we present the time-series up to the latest common year with complete citation data, which is 2017 for the wos_b_2020 and wos_b_2021 databases. This comparison highlights any differences in trends between the databases for the most recent decade.

For other indicators, it is most useful to compare changes between just the most recent years of complete data in each database. For instance, we examine the threshold for Excellence Rates in 2017 from the wos_b_2020 database against 2018 in the wos_b_2021 database. Changes between the years are expected given we are comparing two different sets of publications, however this comparison can also provide insight into structural changes between the database iterations, such as the addition or removal of journals from indices, which may influence indicators at the macro-level.

Such comparisons are also helpful in identifying new or removed institutions or discipline categories. Further, although users will likely use the latest database to produce a complete time-series for new analyses, it is important to understand how additional years of a time-series might differ to existing time-series presented in publications and reports.

Set of entities

We have chosen to compare the databases at the national, sectoral, and institutional levels. The countries chosen are based on those most commonly examined by the DZHW as countries against which it is useful and informative to compare Germany. We also examine the key German sectors: Universities (Uni), Fachhochschulen (FH), Max Planck Gesellschaft (MPG), Fraunhofer Gesellschaft (FHG), Helmholtz Gemeinschaft (HGF), Leibniz Gemeinschaft (WGL), the business sector (Econ), non-university hospitals (Klinik), and combined Ressortforschung-Bund and Ressortforschung-Laender (Gov). The remaining smaller sectors, such as research associations, clubs, and international and foreign organisations are grouped into an "other" category. Individual German institutions are also able to be examined via the KB's institutional coding for Germany. However, as there are a large number of institutions, we present data only for institutions that have shown substantial changes in the indicator of interest.

Methodological details

We focus on articles and reviews published in journals as these are the most common documents used in bibliometric analyses. Unless otherwise stated, we examine content indexed in the Science 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 [2] determined that at least 3 years is required for publications to reach their maximum number of citations per year, after which point the number of citations are likely representative of the publication's long-term impact. As such, citation-based indicators include all citations received within the publication year and the subsequent two years.

Whole counting is used throughout the report. Although it is most common to use fractional counting, analysing variables using whole counts will still reveal potential changes in the variables.

Data for disciplines are presented based on either the sc_traditional, sc_extended or Research Areas (RA) classification. Sc_traditional is the fine-grained classification more commonly used in analyses by the DZHW. 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 present some data on the RA classification. The RA consists of five broad groups derived from a mapping of sc_extended disciplines provided by Clarivate Analytics. Each section containing data about disciplines notes which classification is used.

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, 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 Research Area.

Changes in publication counts over time may reflect changes made by countries, the database provider, and/or administrative decisions. For example, it is expected that the wos_b_2021 database contains a greater number of publications for the most recent years than the wos_b_2020 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.

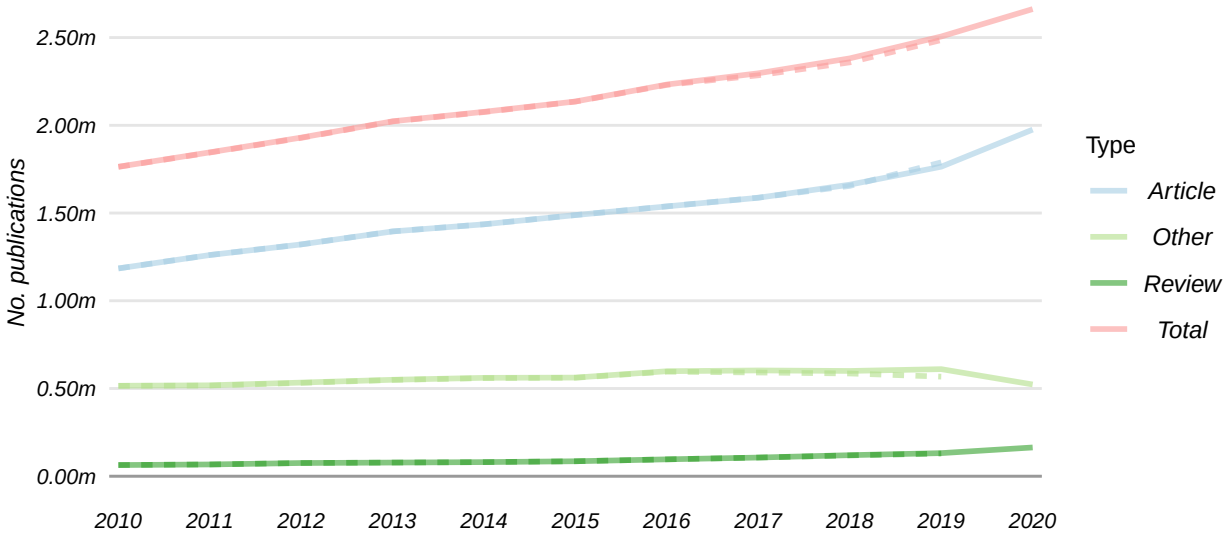

Figure 1: Number of publications by document type and database, where dashed lines show the previous database and full lines show the current database.

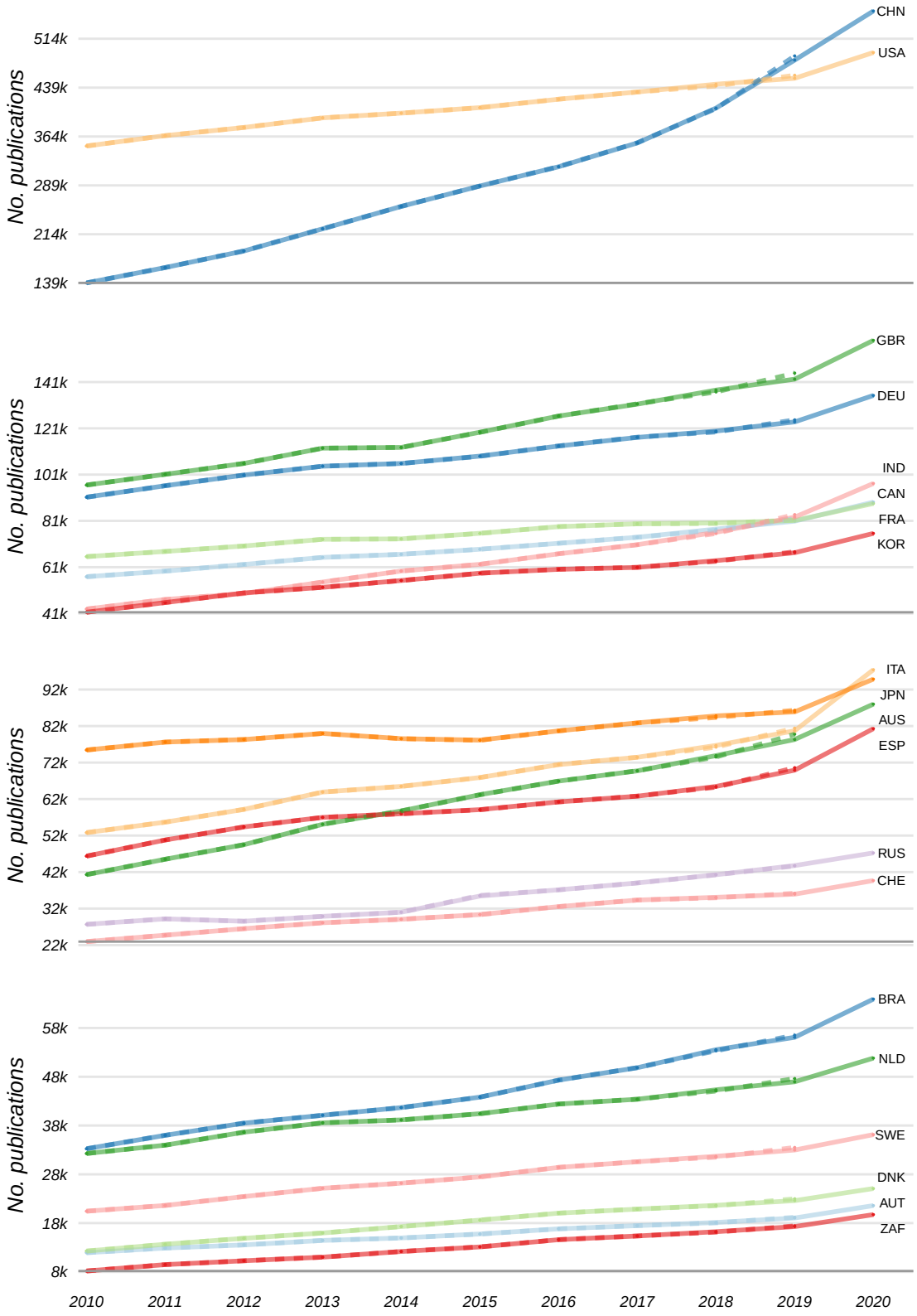

Figure 2: Whole counts of national publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels have different axes.

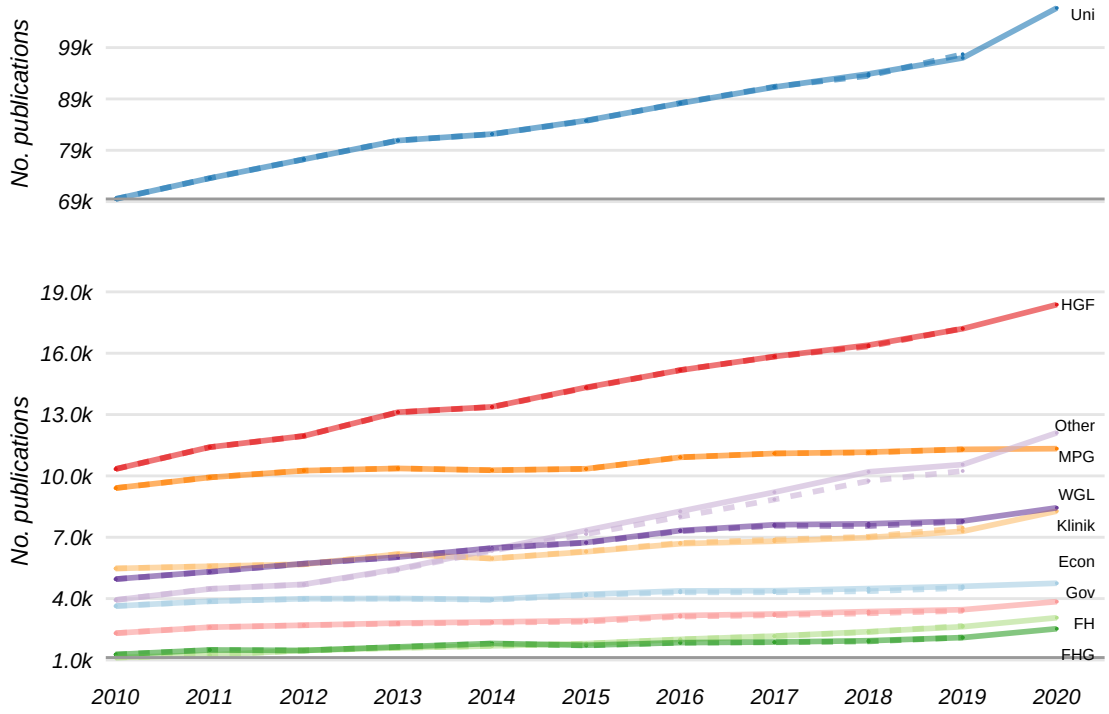

Figure 3: Whole counts of sectoral publications by database, where dashed lines show the previous database and full lines show the current database. Please note the panels' different scales.

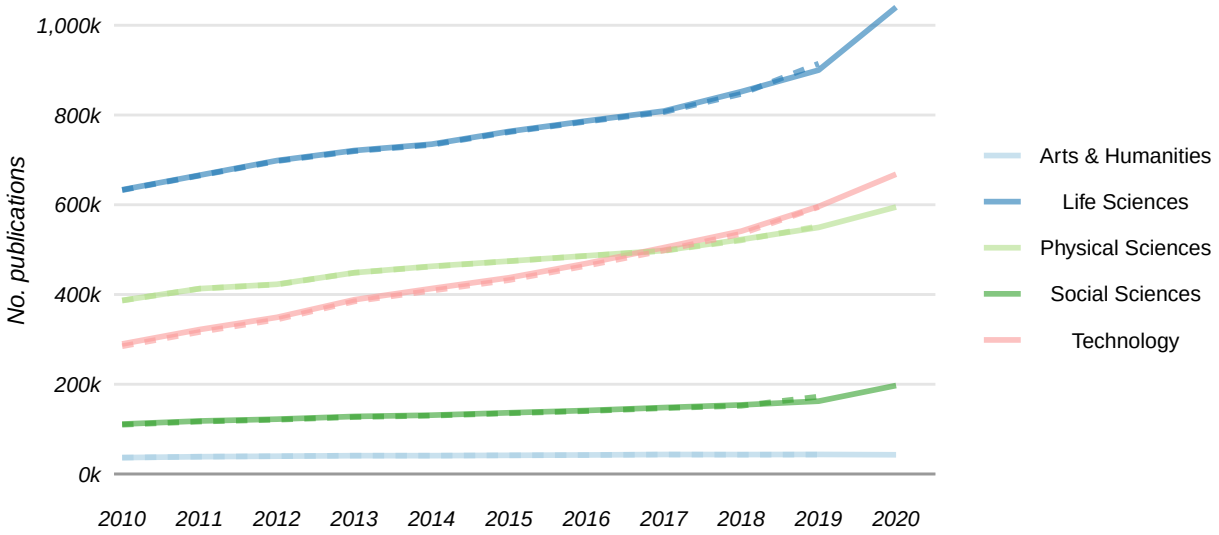

Figure 4: Whole counts of publications by Research Area and database, where dashed lines show the previous database and full lines show the current database.

Journals: Total indexed and numbers added or removed

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

As all sources indexed have titles – as opposed to some missingness of ISSNs – changes in journals were identified by matching the titles of all journals indexed in 2019 in the wos_b_2020 database to those with 2020 content in the wos_b_2021 database. Titles in wos_b_2020 but not wos_b_2021 were considered removed, while titles present in wos_b_2021 but not in wos_b_2020 were considered added. These data may include a small number of journals that changed titles. Journals were mapped to the Research Areas classification from the sc_extended classification. Some double-counting of journals between RAs occurs where the journal is assigned to two or more classifications mapped to different RAs. In total, 279 journals were added and 192 were removed.

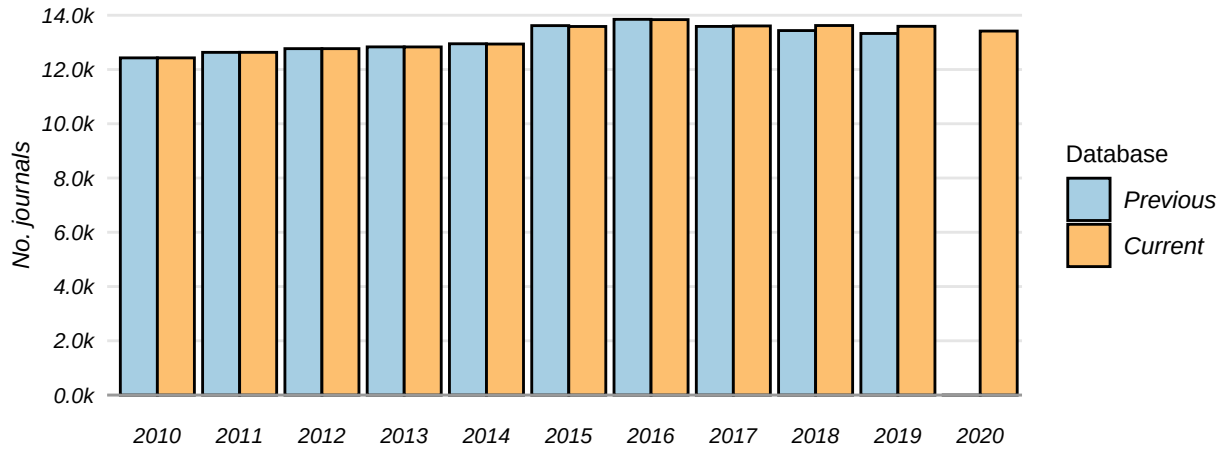

Figure 5: The number of journals indexed in each source over time.

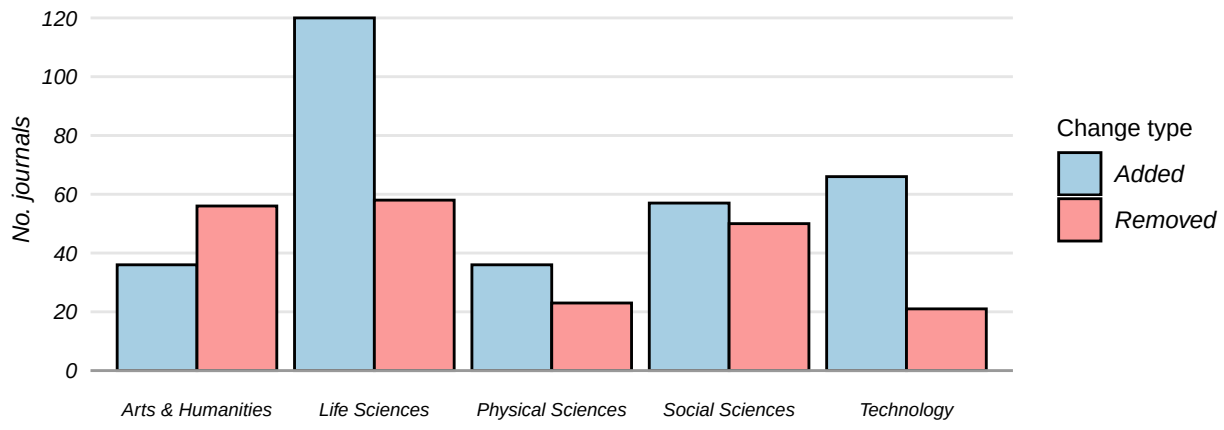

Figure 6: The number of journals added and removed between the latest years in each database by Research Area.

Excellence Rates: Selected countries and German sectors

Excellence Rates (ER) identify the percentage of an entity's publications that are in the 10% most highly cited publications from each discipline and could be considered of excellent quality on this basis. ERs are a common indicator used to assess an entity's performance, with an ER exceeding the expected 10% threshold interpreted as better than expected performance. ERs are calculated here based on the sc_traditional discipline classification. ERs for the most recent years from the two databases are presented for German sectors in Figure 7 and for countries in Figure 8. As with whole counts of publications, we would expect general agreement between the databases, particularly in the earlier years of the time-series, so substantial deviations may warrant further analysis.

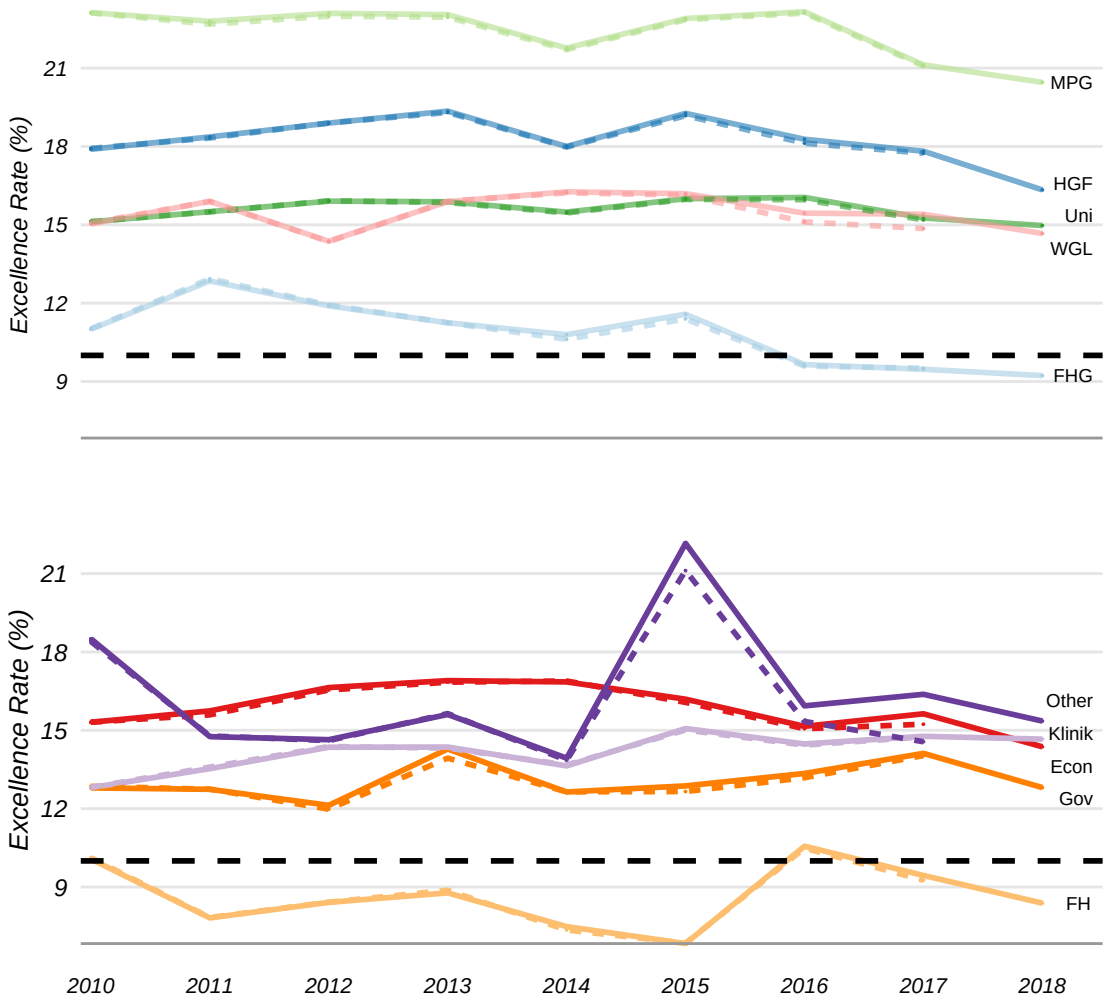

Figure 7: Excellence rates by sector, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

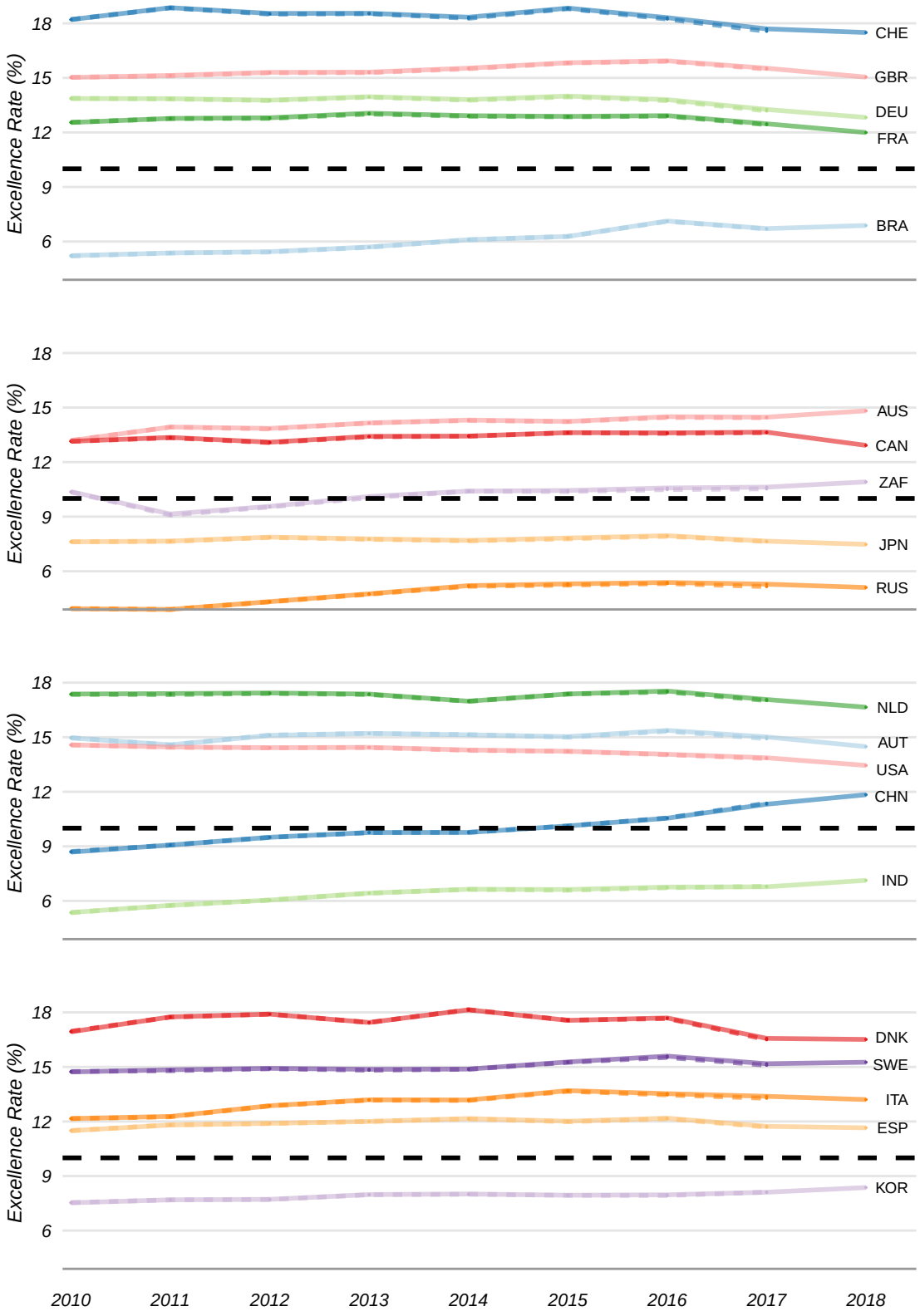

Figure 8: Excellence rates for selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database. The black line is the expected 10% threshold.

Excellence Rates: Thresholds by discipline

ERs are dependent on the number of citations a publication receives in relation to the threshold it must exceed to reach the top 10% of the pool of reference publications. A change in the 10% threshold for a discipline can make it more or less difficult for a publication to exceed the threshold, which can have knock-on effects for a sector or country's ER over time. For example, substantial differences in countries' ERs between WoS and Scopus were observed in Stahlschmidt, Stephen and Hinze [1]. This results from differences in coverage between the two databases, as Scopus' greater coverage of more sparsely cited journals lowers the ER threshold and allows high-performing countries to receive higher ERs. The greater consistency of coverage in WoS, compared to between WoS and Scopus, means we expect less change in the ER thresholds between the iterations of the WoS databases. However, changes in the journals indexed may influence the ER threshold for disciplines, potentially affecting the ERs of countries or, in particular, sectors due to their stronger disciplinary focus.

To examine changes in thresholds, we present in Figure 9 the ER thresholds for articles and reviews in each sc_traditional discipline. We assess articles and reviews separately given the known differences in citation patterns between the document types. Large increases in the threshold would require publications to achieve substantially more citations to exceed the 10% threshold and be included in the ER, while a decrease in the threshold means publications require fewer citations than previously.

In the top panels of Figure 9 we see the ER thresholds for each discipline in 2017 in both the wos_b_2020 and wos_b_2021 databases. The colour denotes the number of disciplines with each combination of thresholds, from fewer in blue to more in red. These panels depict the changes in ER thresholds in the same year between databases, providing context for any differences observed in 2017 in Figures 7 and 8. In the bottom panels we present again the thresholds for each discipline in 2017 in the wos_b_2020 database but now compared against the threshold in 2018 in the wos_b_2021 database. These panels highlight changes between the latest years in each database, indicating whether we could expect to see changes in ERs between the databases.

The outlying disciplines with the greatest change in thresholds in the bottom panels of Figure 9 are shown in Tables 1 and 2, along with disciplines where the previous threshold was zero, highlighting potentially new or emerging disciplines.

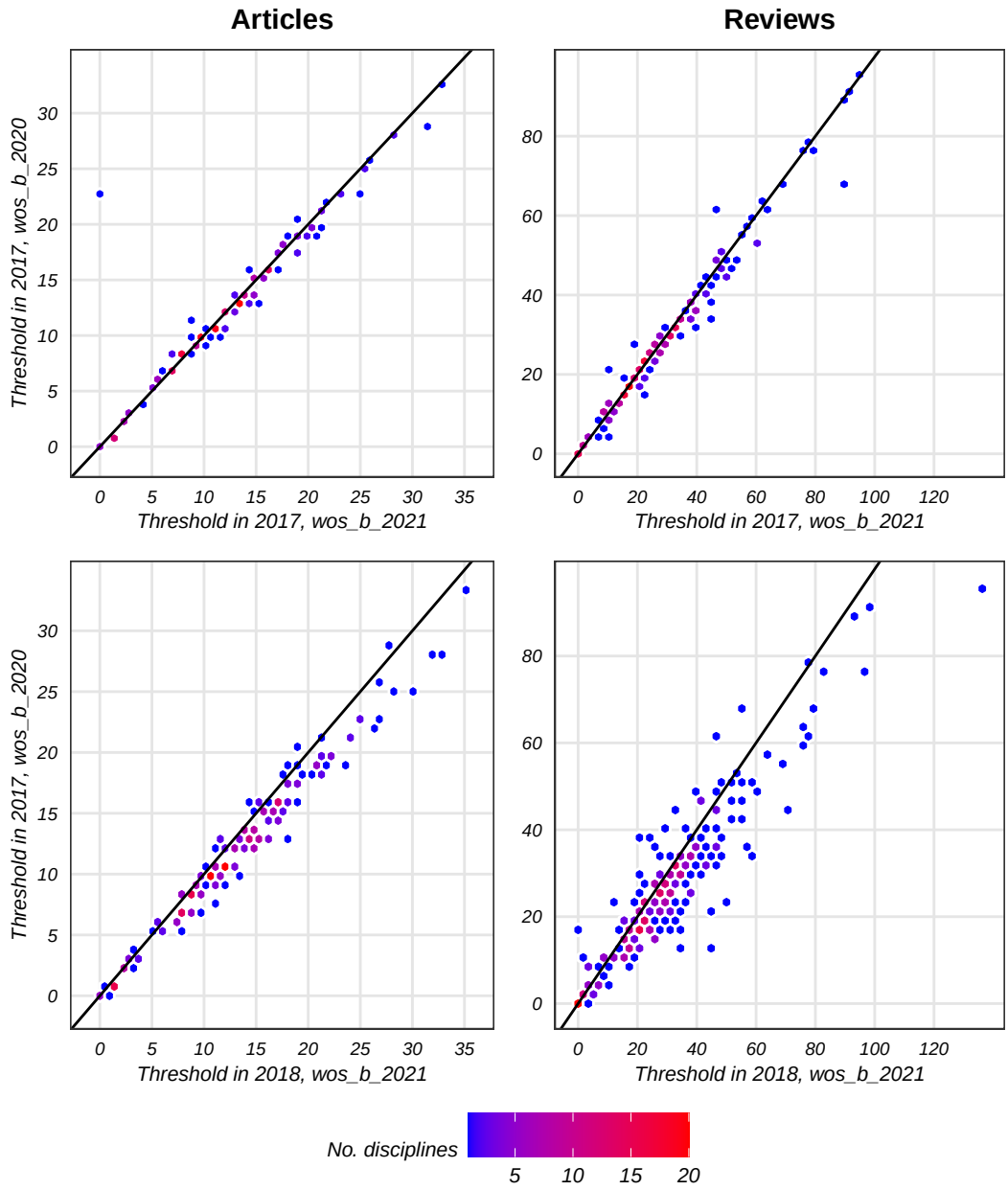

Figure 9: The ER threshold for articles and reviews in each discipline (sc_traditional) between databases, where colour denotes the number of disciplines with this combination of thresholds.

Table 1: Articles: Disciplines where the ER threshold decreased, or increased by over 40% between 2017 in wos_b_2020 and 2018 in wos_b_2021, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Dance	0	1	301	Inf
Ethnic Studies	5	8	967	60.0
Psychology, Psychoanalysis	2	3	471	50.0
Ethics	7	10	2,481	42.9
Computer Science, Cybernetics	29	28	1,517	-3.4
Critical Care Medicine	20	19	4,310	-5.0
Neuroimaging	19	18	2,931	-5.3
Respiratory System	16	15	8,829	-6.2
Rheumatology	16	15	4,101	-6.2
Psychology, Biological	13	12	1,772	-7.7
Transplantation	13	12	4,010	-7.7
Geology	12	11	2,950	-8.3
Ophthalmology	11	10	8,632	-9.1
Quantum Science & Technology	16	14	2,419	-12.5
Logic	4	3	909	-25.0
Literary Theory & Criticism	1	0	846	-100.0

Table 2: Reviews: Disciplines with a current or previous ER threshold of at least 5, where the threshold decreased by over 25%, increased by over 60% between 2017 in wos_b_2020 and 2018 in wos_b_2021, or the previous threshold was 0.

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Asian Studies	0	2	30	Inf
Classics	0	1	29	Inf
Public Administration	12	44	28	266.7
Engineering, Petroleum	12	33	68	175.0
Geology	16	34	89	112.5
Medicine, Legal	8	17	124	112.5
Neuroimaging	24	50	153	108.3
Ethnic Studies	5	10	14	100.0
Mathematical & Computational Biology	22	44	170	100.0
Engineering, Geological	16	30	51	87.5
Engineering, Marine	17	29	39	70.6
Paleontology	10	17	79	70.0

Discipline	Previous threshold	Current threshold	No. crrnt pubs.	Perc. diff
Anatomy & Morphology	15	25	158	66.7
Geography	15	25	89	66.7
Physics, Particles & Fields	35	58	174	65.7
Agriculture, Multidisciplinary	20	33	297	65.0
Mining & Mineral Processing	16	26	117	62.5
Family Studies	13	21	138	61.5
Psychology, Mathematical	23	37	48	60.9
Mathematics, Applied	12	19	68	58.3
Engineering, Ocean	19	30	27	57.9
Optics	45	70	417	55.6
Psychology, Social	22	34	70	54.5
Transportation	25	38	68	52.0
Engineering, Manufacturing	37	56	175	51.4
Microscopy	37	27	83	-27.0
Medical Ethics	11	8	32	-27.3
Materials Science, Coatings & Films	40	29	118	-27.5
Automation & Control Systems	45	32	104	-28.9
Marine & Freshwater Biology	30	21	331	-30.0
Robotics	39	24	52	-38.5
Social Issues	23	13	60	-43.5
Materials Science, Paper & Wood	38	20	38	-47.4
Demography	9	4	13	-55.6
Archaeology	8	3	142	-62.5
Cultural Studies	10	1	47	-90.0
Logic	17	1	2	-94.1

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

The number of citations a publication could be expected to receive is dependent to an extent on its discipline. As such, we examine here the mean 3-year citations of articles and reviews by discipline. Mean 3-year citations (MC3) are the mean citations publications in each discipline accrued in the first 3 years after publication. As we did with ERs, we examine here in Figure 10 the last common year in both databases (top panels) to assess the retroactive effects stemming from changes made in the latest database, and the latest complete year in both databases (bottom panels) to assess potential structural changes and updates to the time-series. A greater deviation of disciplines from the central line indicates a greater degree of change in the mean citations of a discipline's items between years. Data are based on the sc_traditional discipline classification. The outlying disciplines from the bottom panels of Figure 10 are shown in Tables 3 and 4, along with disciplines where the previous threshold was zero. We use a threshold of a current MC3 of at least 1 for articles and 3 for reviews to remove disciplines with spurious changes due to low levels of citations.

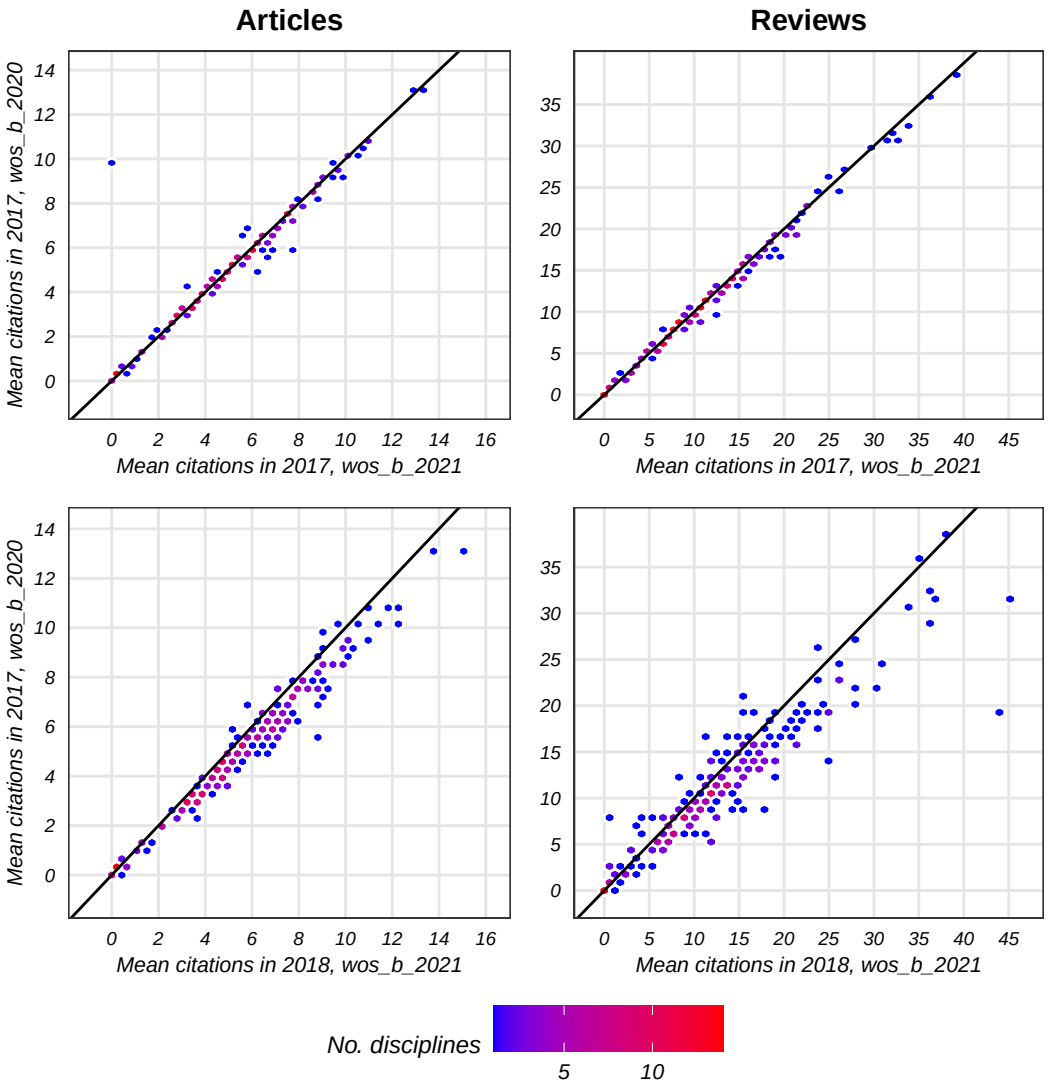

Figure 10: The MC3 for articles and reviews in each discipline between databases, where colour denotes the number of disciplines with this combination of citations.

Table 3: Articles: Disciplines with a current MC3 of at least 1, where the MC3 decreased by over 20% or increased by over 50% between 2017 in wos_b_2020 and 2018 in wos_b_2021, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs.	Perc. diff.
Ethnic Studies	2.2	3.5	967	64.80
Psychology, Mathematical	5.5	8.8	685	60.75
Communication	3.4	5.1	4,242	47.02
Hospitality, Leisure, Sport & Tourism	5.0	6.8	3,230	37.06
Philosophy	0.8	1.1	6,163	33.19
Engineering, Petroleum	3.6	4.8	2,555	32.02
Medical Informatics	5.2	6.8	3,707	31.94

Table 4: Reviews: Disciplines with a current MC3 of at least 3, where the MC3 decreased by over 20% or increased by over 60% between 2017 in wos_b_2020 and 2018 in wos_b_2021, or the previous MC3 was 0.

Discipline	Previous cit.	Current cit.	No. crrnt pubs.	Perc. diff.
Public Administration	5.0	12.2	28	142.38
Physics, Particles & Fields	19.1	43.8	174	129.88
Horticulture	5.6	11.5	116	106.29
Mathematical & Computational Biology	9.1	18.2	170	99.62
Language & Linguistics	2.7	5.3	38	99.19
Engineering, Petroleum	5.8	11.1	68	89.97
Limnology	8.6	15.5	22	80.22
Astronomy & Astrophysics	13.8	24.5	284	78.21
Engineering, Marine	5.7	9.9	39	72.55
Mathematics	1.9	3.1	32	69.08
Engineering, Ocean	8.6	14.4	27	66.93
Neuroimaging	11.8	19.1	153	61.27
Rheumatology	14.2	11.3	855	-20.09
Automation & Control Systems	16.8	13.3	104	-20.89
Women's Studies	4.3	3.1	41	-27.32
Materials Science, Coatings & Films	20.9	15.1	118	-27.68
Psychology, Educational	12.0	8.4	110	-29.44
Acoustics	16.5	11.6	134	-29.79
Ethics	6.4	4.3	80	-33.27
Statistics & Probability	7.7	5.0	102	-34.89
Social Issues	7.8	4.3	60	-44.56
Medical Ethics	6.7	3.1	32	-53.21

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

While ERs represent the most highly cited publications and mean citations tell us about what's average, the percentage of uncited publications can tell us about the entities at the tail end of the citation distribution. When examining uncited publications, we expect to see a decreasing trend in uncited publications over time. This occurs because citation counts are based on the items indexed in each database and so, as Clarivate Analytics continues to index journals, the likelihood increases that any publication will have been cited by the indexed items. In particular, we would expect that the percentage of uncited publications in the last common year would be lower in the current database than the previous database, as data added in the latest iteration "complete" the incomplete last year of the previous database. An increase in uncited publications in the latest year may reflect processing issues that require investigation. We present in Figures 11 and 12 the percentage of articles and reviews per German sector and selected country that remained uncited 3 years after they were published.

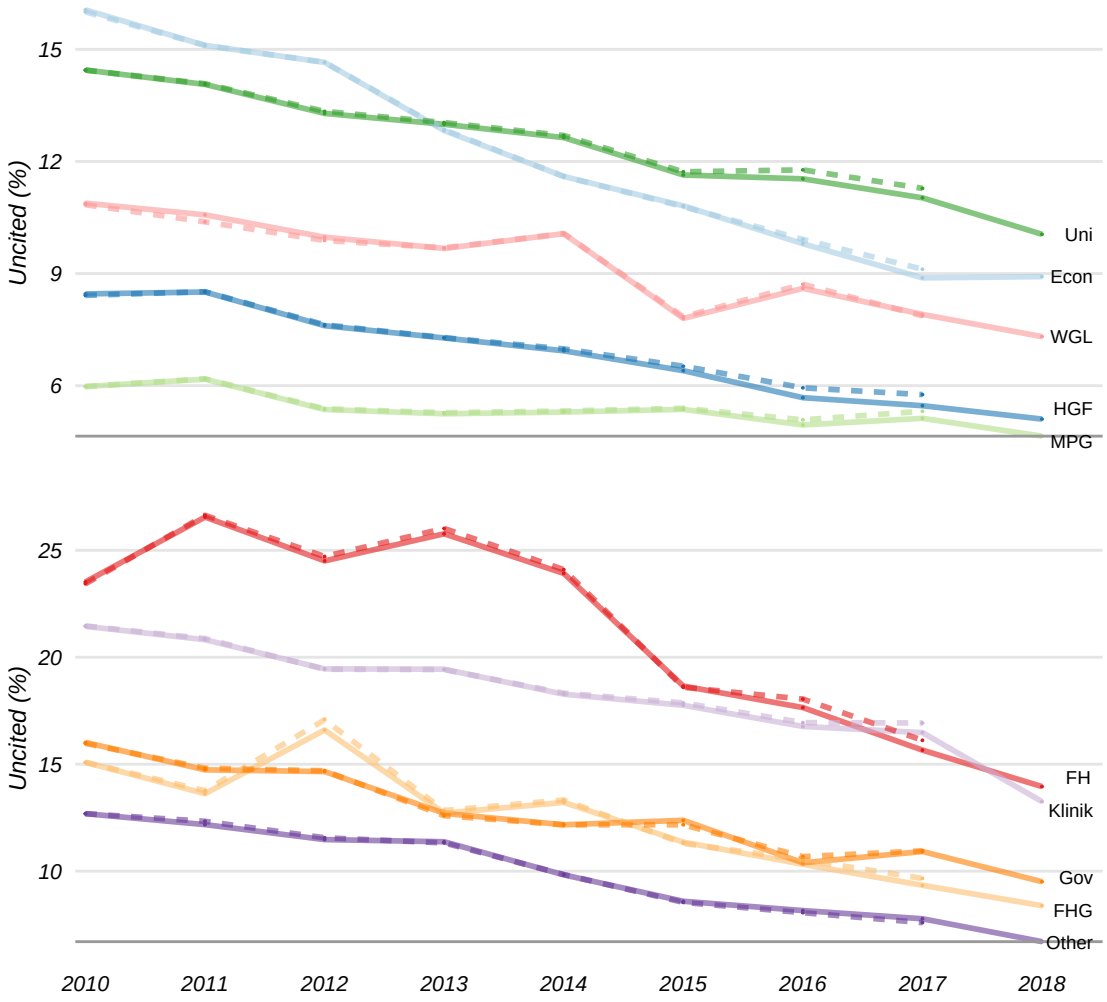

Figure 11: The percentage of uncited publications by German sector, based on whole counts, where dashed lines show the previous database and full lines show the current database.

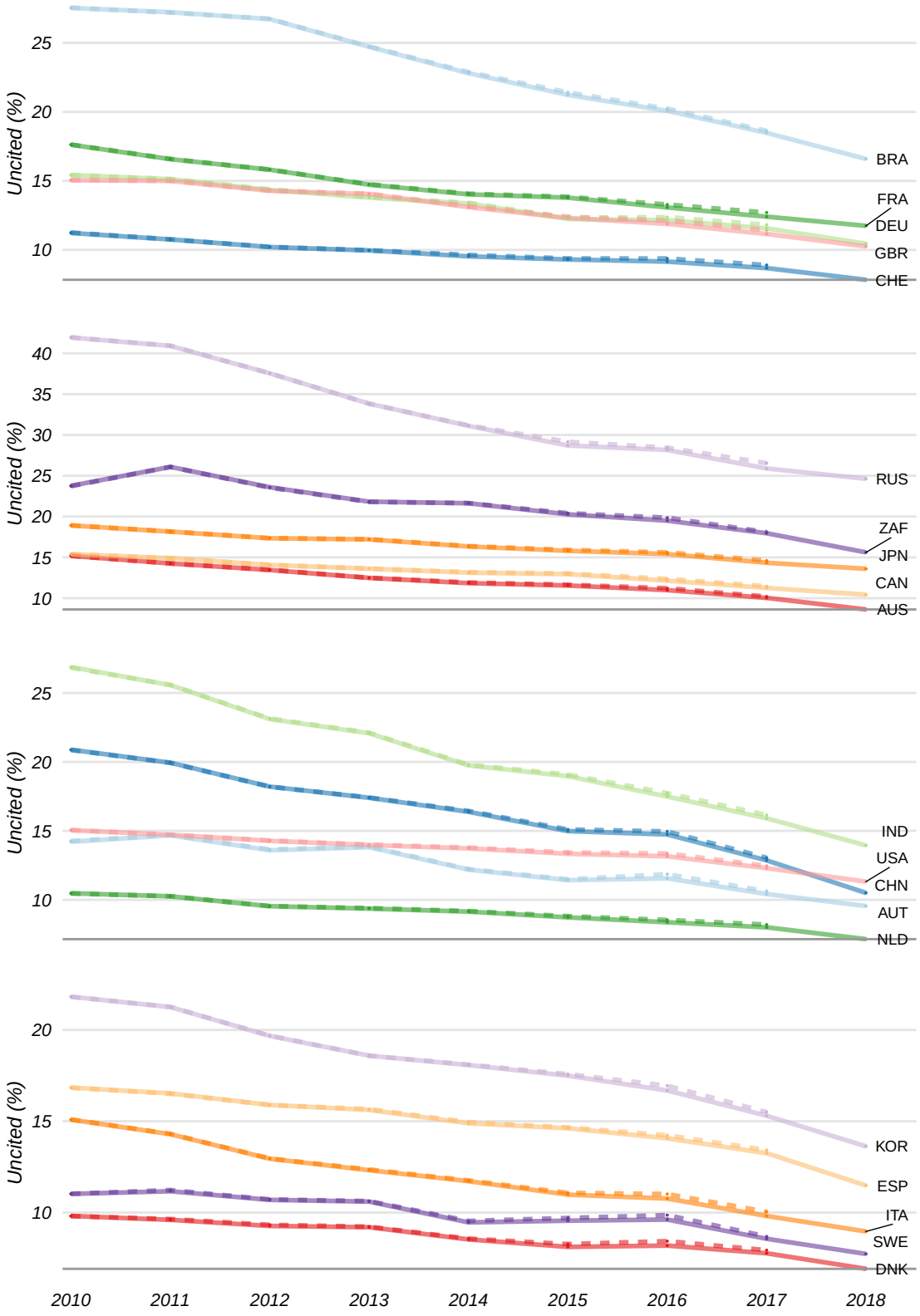

Figure 12: The percentage of uncited publications by selected countries, based on whole counts, where dashed lines show the previous database and full lines show the current database.

Disciplines: Changes in discipline classification

This section highlights any changes that have been made to WoS' sc_traditional discipline classification in Table 5 and the sc_extended classification in Table 6. 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. Figure 13 shows the number of publications assigned to specific sc_traditional disciplines identified to have changed in recent versions of the database.

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

Classification	Previous pubs	Current pubs
GREEN & SUSTAINABLE SCIENCE & TECHNOLOGY	495	NA
Planning & Development	126	NA

Table 6: Changes in the sc_extended discipline classification structure between the previous and current databases.

Classification	Previous pubs	Current pubs

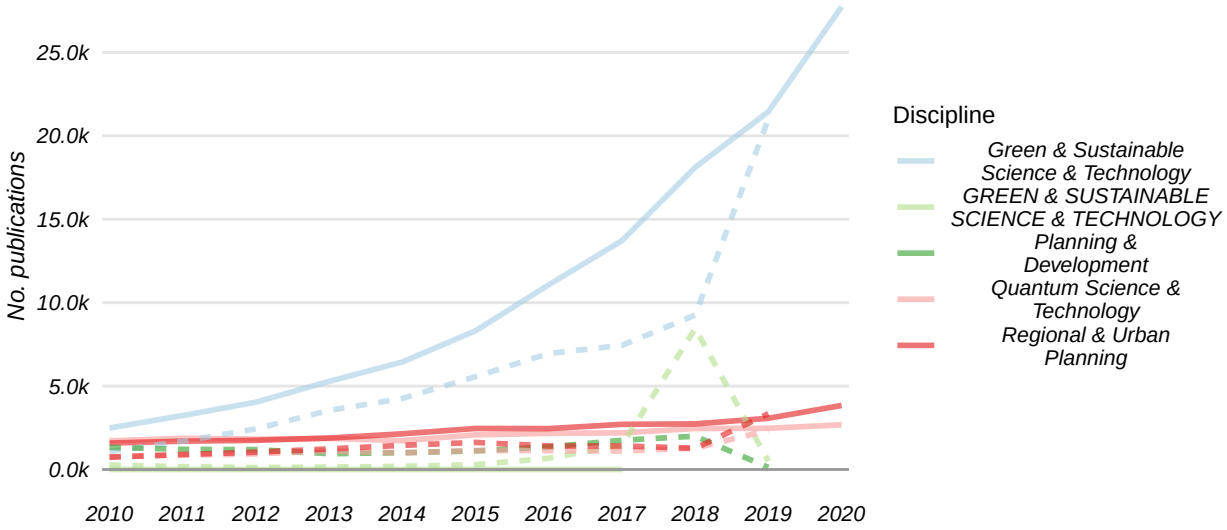

Figure 13: Time-series of sc_traditional disciplines known to have changed in recent versions of the databases. Dashed lines show the previous database and full lines show the current database.

Disciplines: Changes in articles and reviews by discipline

This section identifies the disciplines that had a substantial change in the number of publications assigned to them between the latest years in each database. Changes in counts of publications per discipline reflect changes in the journals indexed, the classification structure, and any potential processing issues. As such, any large changes shown here may be worth examining.

We show in Figure 14 the 20 disciplines with the highest percentage increases and decreases in publication counts between 2019 in wos_b_2020 and 2020 in wos_b_2021. The number shown next to each bar is the numerical change in publication counts. We have used whole counting and the disciplines are based on the sc_traditional classification. Disciplines previously identified as being new or removed have not been included here.

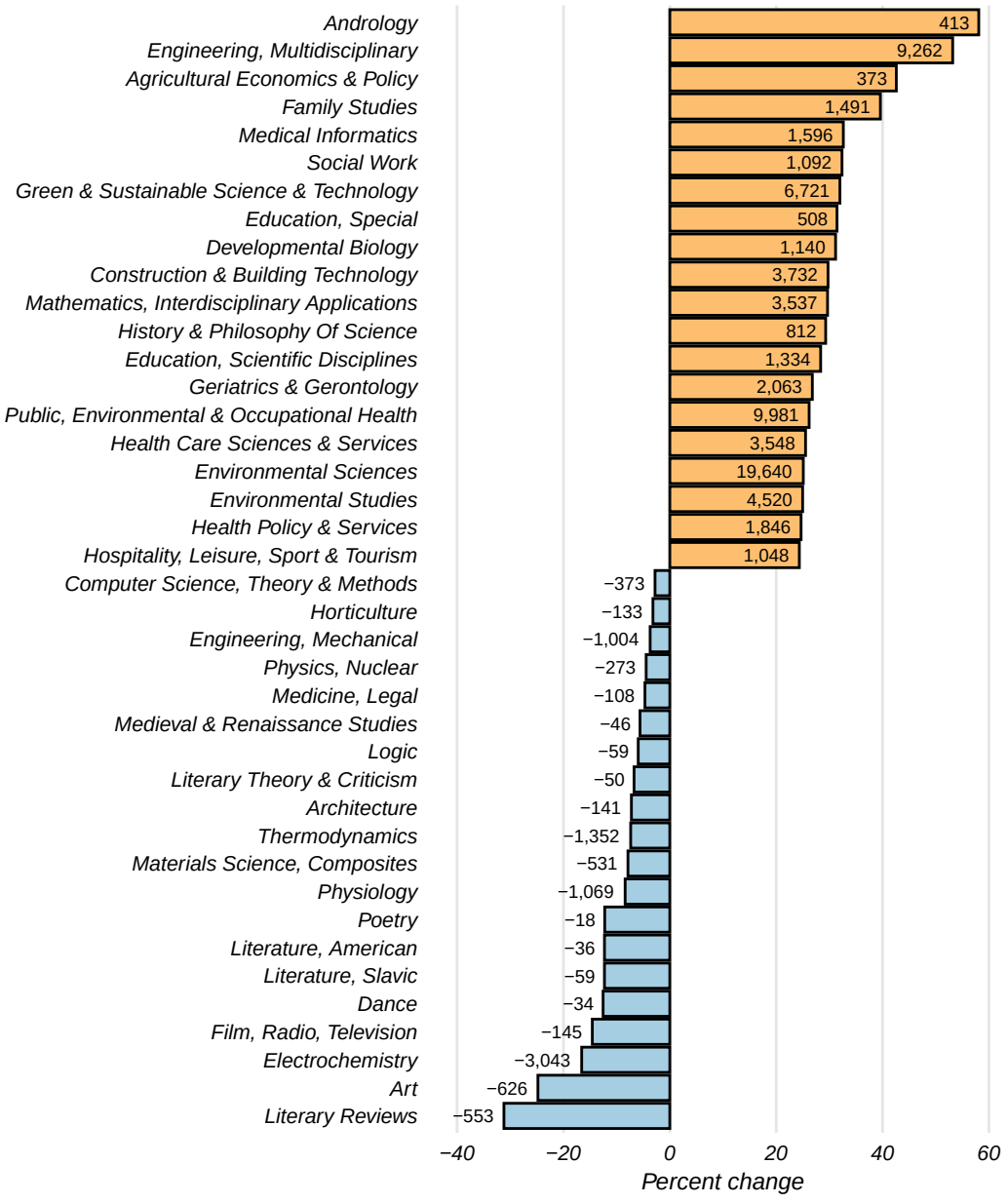

Figure 14: The 40 disciplines with the highest percentage change in publication counts between 2019 in wos_b_2020 and 2020 in wos_b_2021, with numerical difference in counts.

Disciplines: Number of publications not assigned to a discipline

This section presents in Figure 15 the percentage of publications in each database that were not assigned to a discipline over the previous 10 years. Complete assignment of publications to disciplines is important as citation-based indicators typically use field-normalisation to account for differences in citation practices between disciplines. As such, items missing discipline information are excluded from such analyses and so large percentages of, or large changes in, unclassified items should be investigated.

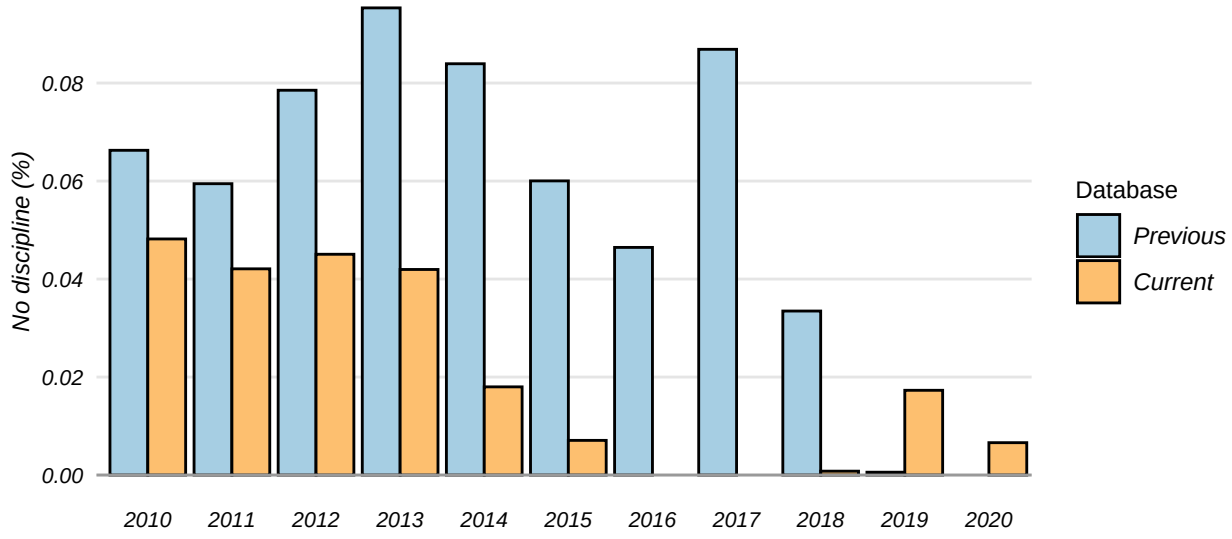

Figure 15: The percentage of publications in each database that do not have a discipline classification.

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

This section details the number of items for which changes were made to key metadata in the latest iteration of the database or the items were removed. We look at changes in the recorded publication year, document type and publication type as these three variables are typically the key inclusion criteria for bibliometric analyses. We also examine the number of items that were present in the wos_b_2020 database but not in the wos_b_2021 database. A change in metadata for a large number of items may be problematic, particularly if the changes are not randomly distributed, such as adjustments having been made to items from a particular journal or set of publications, which may affect counts and indicators for specific entities. Some changes can be expected as Clarivate Analytics updates or corrects items, however changing or removing a large number of items may require investigation.

We identify changes in the metadata of in-scope items by first matching items between the wos_b_2020 and wos_b_2021 databases using the UT_EID identifier and then counting the number of instances where matched items do not have the same publication year, document type (i.e. an article or review has been changed to a different document type) or publication type (i.e. the publication type changed from journal to another type) between databases. As such, Table 7 shows the number of items that have had their metadata changed between the previous and current databases. The number of items removed is based on UT_EIDs in wos_b_2020 not matching a record in wos_b_2021. Data are presented based on the publication year recorded in the wos_b_2020 database.

Table 7: The number of items with changes in metadata or removed between the previous and current database versions.

Year	Pub. year	Doc. type	Pub. type	Removed
2010	8	1	0	116
2011	0	3	0	190
2012	0	4	0	255
2013	1	5	0	311
2014	11	7	0	527
2015	23	6	0	847
2016	67	25	0	1,642
2017	129	403	0	2,278
2018	140	972	0	1,543
2019	51,115	8,129	0	53,908

Metadata: Publications from incorrect 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 inclusion of items from other indices, such as the Emerging Sources Citation Index, can be problematic as these items may fundamentally differ from those in the three core indices in, for instance, the countries of their authors and publishing journals, which can influence citation-based indicators. As such, we check here whether the database includes items indexed outside of the SSCI, SCIE and AHCI. Tables 8 and 9 show the annual number of items from outside these indices present in the wos_b_2020 and wos_b_2021 databases respectively. Blank tables indicate there are no incorrect inclusions.

Table 8: Number of publications indexed in out of scope indices, wos_b_2020.

PUBYEAR

Table 9: Number of publications indexed in out of scope indices, wos_b_2021.

PUBYEAR

Metadata: Missing metadata variables

Figure 16 shows the annual percentage of publications in each database that are missing particular metadata, including page numbers, journal issue and volume information, DOIs, titles, references, abstracts, and keywords. We could reasonably expect improvements over time in missing metadata, such as for DOIs through increasing uptake of this identifier, however increasing missing metadata should be investigated. Empty graphs indicate there were no items missing this metadata.

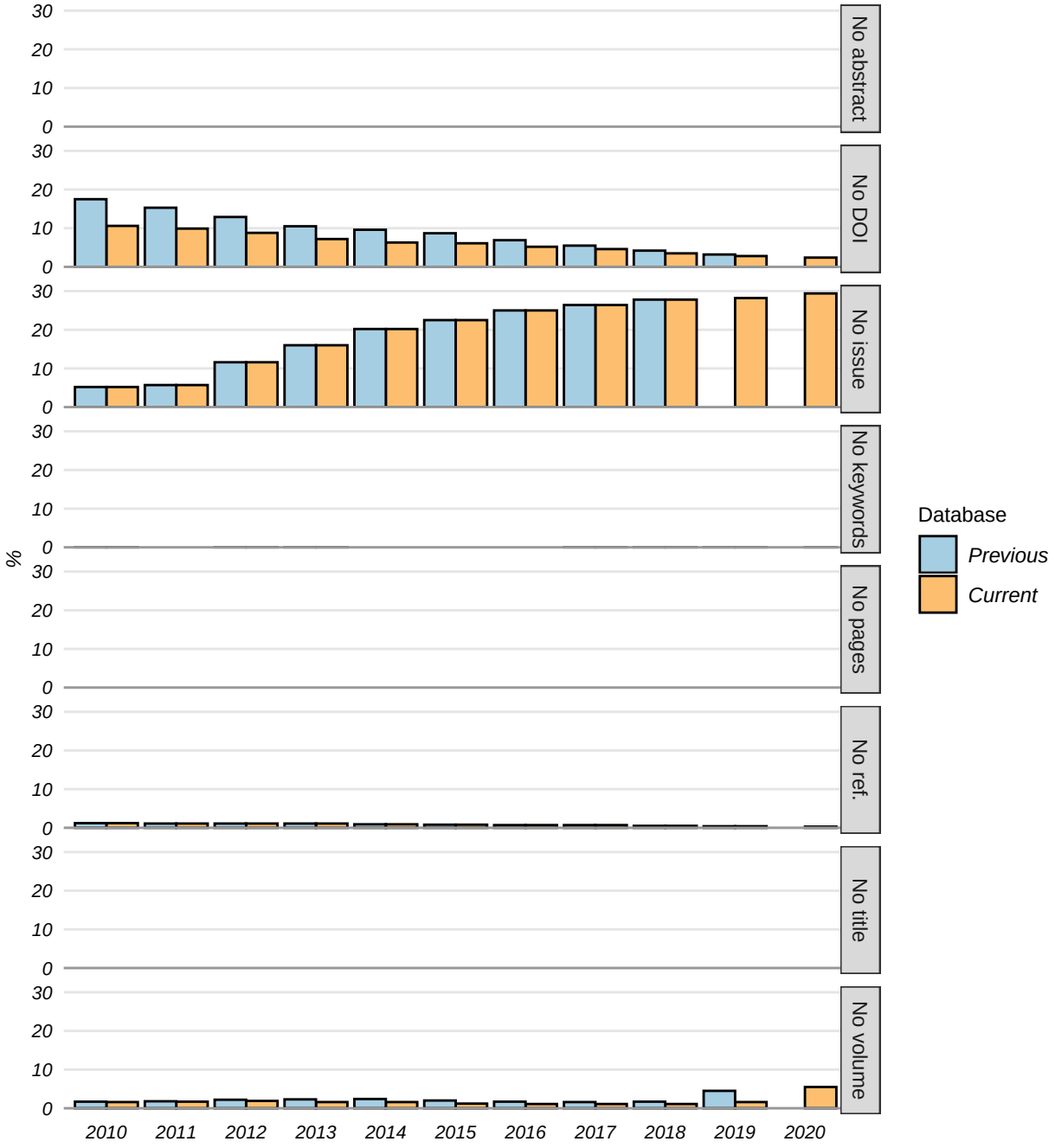

Figure 16: The percentage of items with missing metadata over time by database.

Institution and country data: Number of articles and reviews with missing data

Bibliometric analyses often examine indicators at the level of institutions or countries. Further, fractional counting can be applied based on institutions, with articles apportioned according to authors' affiliations. It is imperative for accurate indicators that most, if not all, items have institution and country data, as missing information removes otherwise valid items from analyses.

The Items table of the KB databases holds a record of all available items, while the associated data about authors' affiliations are held, in part, in the Institutions table. We have operationalised missing institution information here as publications that appear in the Items table but have no corresponding information in the Institutions table. We present in the top panel of Figure 17 the number of items in each database between 2010 and 2019 with no institution information. Additionally, items can have institution information but no country code – from which country counts are derived – and these are shown in the bottom panel of Figure 17. Large disparities between the databases or substantial increases in missing information should be investigated.

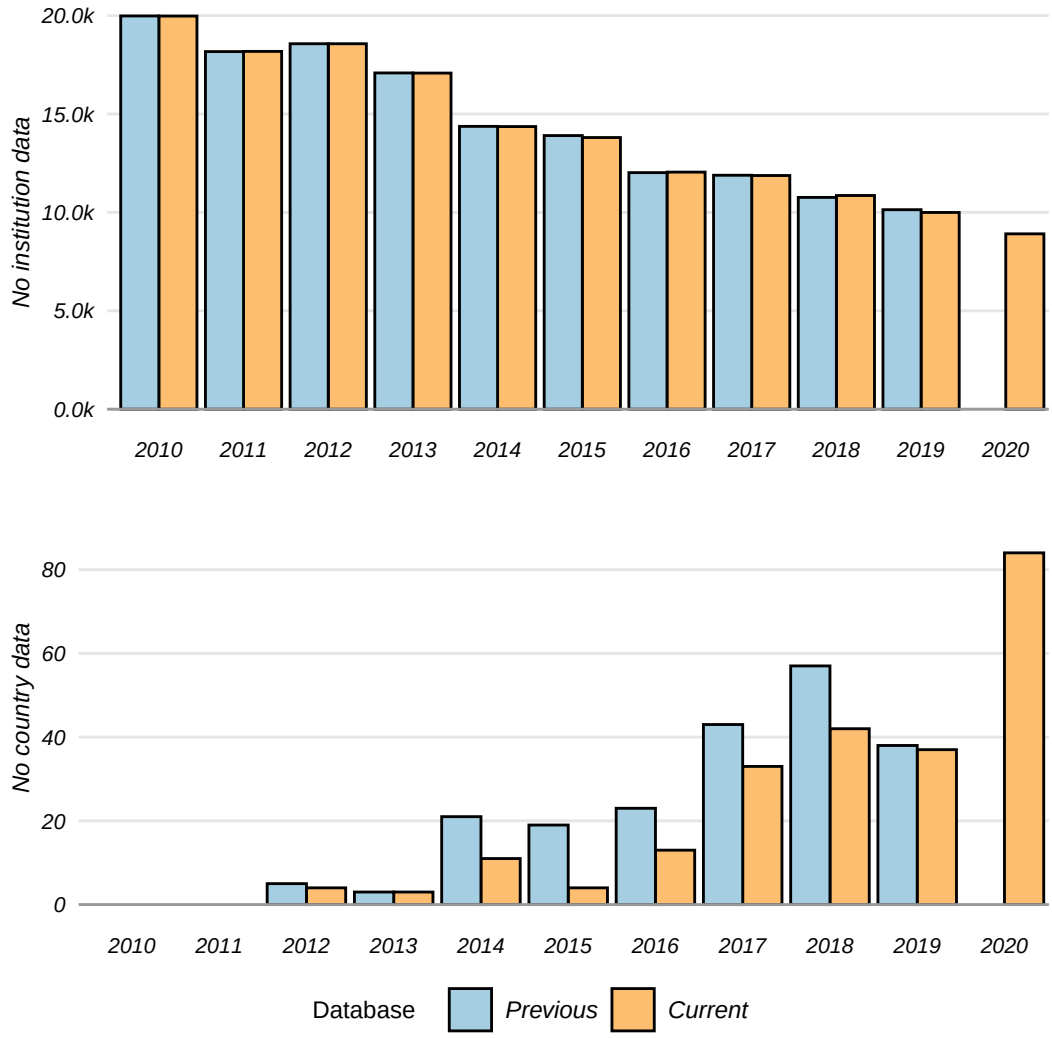

Figure 17: The number of items with missing institution information (top) and the additional items that have institution information but no country code (bottom) over time by database.

Author-institution links: Percentage complete by 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_extended discipline with complete links between authors and institutions.

In Figure 18, we see in the left panel the percentage of complete links for 2019 data in both the previous and current databases, highlighting any retroactive changes that may have been made in the current database. In the right panel is again the percentage of complete links made in 2019 in the wos_b_2020, now compared with the 2020 in the wos_b_2021, indicating potential changes between the latest year in each database.

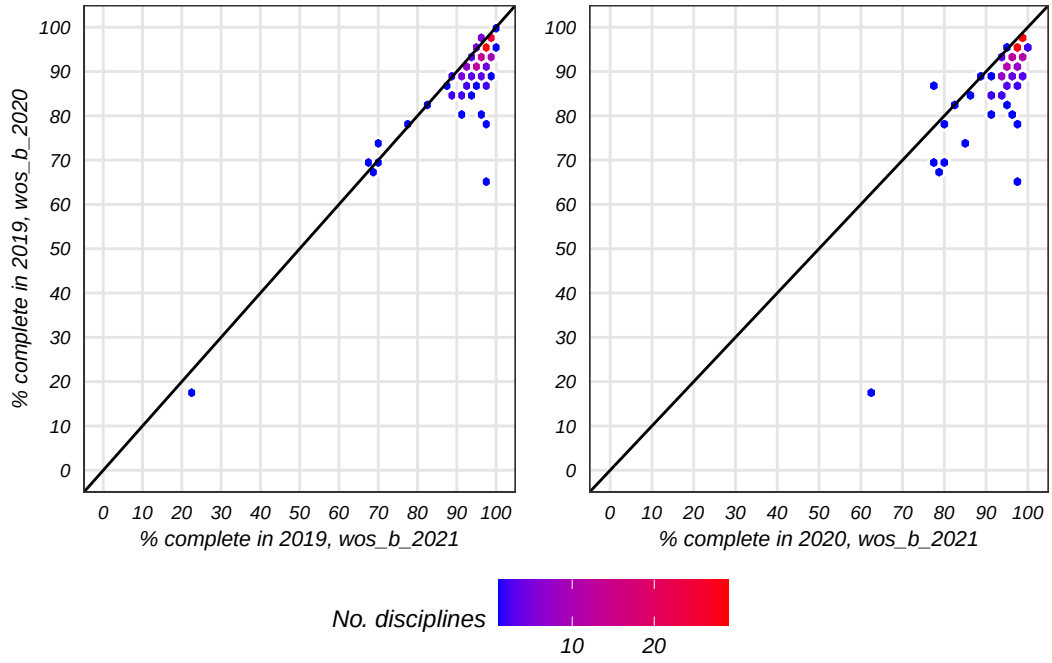

Figure 18: The percentage of complete author-institution links by disciplines (sc_extended).

The outlying disciplines observed in the right panel of Figure 18 that have changed by more than 7 percentage points in the percentage of complete author-institution links between databases are shown in Table 10.

Table 10: Disciplines (sc_extended) that changed by more than 7 percentage points in missing links between 2019 in wos_b_2020 and 2020 in wos_b_2021.

Discipline	Previous items	% prvs complete	Currrnt items	% crrnt complete	Change
Art	61,942	86.5	25,492	78.4	-8.1
Transplantation	137,840	86.0	183,227	93.7	7.7
Mycology	26,216	89.8	73,400	97.8	8.0
Hematology	258,809	85.4	352,814	94.3	8.8
Crystallography	341,724	90.1	383,697	99.0	9.0
Criminology & Penology	82,127	87.3	285,767	96.6	9.3
Chemistry	16,099,579	89.0	25,490,066	98.4	9.4
Biomedical Social Sciences	92,317	86.0	164,316	95.7	9.6
Nursing	203,577	80.7	288,522	90.6	9.9
Allergy	82,156	83.5	127,871	94.3	10.9
Literature	116,117	66.9	101,526	78.2	11.3
Materials Science	15,199,281	87.2	18,061,968	98.5	11.3
Music	14,513	68.7	19,250	80.0	11.3
Film, Radio & Television	11,101	74.1	9,037	85.8	11.7
Rehabilitation	338,468	80.9	532,565	96.3	15.4
Evolutionary Biology	174,478	77.4	176,218	97.2	19.7
Life Sciences & Biomedicine - Other Topics	2,909,116	65.7	4,537,914	98.0	32.2
Dance	2,308	18.2	2,068	62.9	44.7

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 2020 in wos_b_2021. Substantial changes between years or differences between the databases may require investigation of the cause.

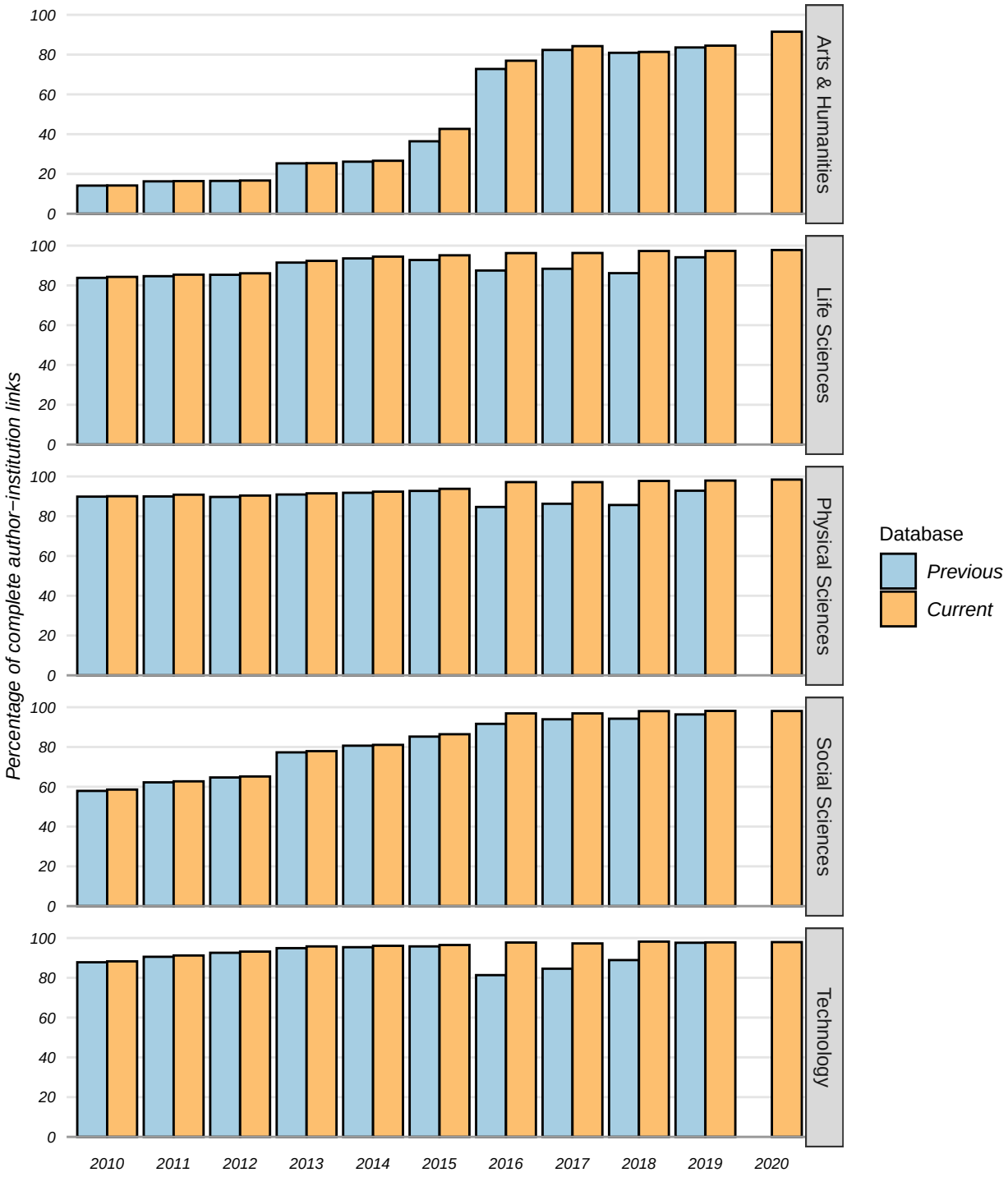

Figure 19: The annual percentage of complete author-institution links by Research Area and database.

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, that were not assigned a KB institution code through the I-Kodierung process. Increases over time may be due to the foundation of new institutions that have not yet been integrated into the coding process. However, publications without KB institutions are typically excluded from sector-level analyses, so it is important to understand the extent of missing institution information.

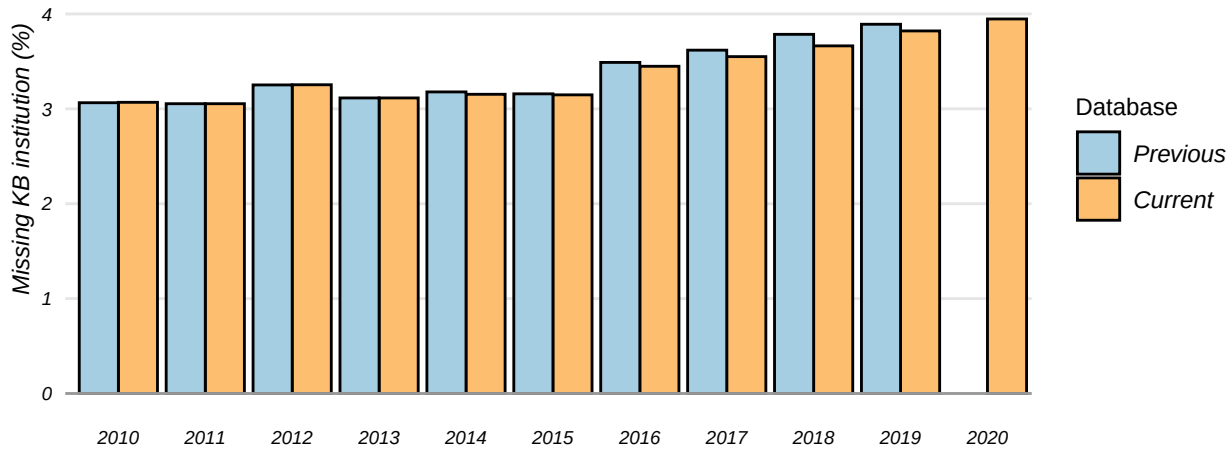

Figure 20: The number of German publications in each database that are missing a KB institution.

German institutions: Changes in whole counts of articles and reviews

This section compares changes in the number of articles and reviews published by German institutions between the latest years available in each database. These tables can assist in identifying institutions for which substantial numbers of publications have been added, removed or otherwise changed in the latest database. They can also aid in assessing the degree of change in publication numbers for larger institutions, which may require further examination if considered unusual or excessive.

Table 11 presents potentially new institutions – these had no publications in 2019 in the wos_b_2020 database but more than five publications in 2020 in the wos_b_2021 database. Conversely, Table 12 shows the institutions that had at least five publications in 2019 in the wos_b_2020 database but no publications recorded in 2020 in the wos_b_2021 database. We also highlight in Tables 13 and 14 the larger institutions (with at least 20 publications) that had a change in publication counts of more than 40% between 2019 and 2020 in the wos_b_2020 and wos_b_2021 databases.

Table 11: Institutions with more than 5 publications in 2020 in wos_b_2021 that had no publications in 2019 in the wos_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs
5,555	Paracelsus Medizinische Privatuniversitat (PMU)	0	147

PK_KB_INST	Name	Previous pubs	Current pubs
11	Senckenberg Gesellschaft fur Naturforschung (SGN)	0	137
5,550	Institute for Advanced Sustainability Studies e.V. (IASS)	0	40
5,529	HHL Leipzig Graduate School of Management	0	28
5,543	Climate Analytics gGmbH	0	24
5,544	Monasterium Laboratory - Skin & Hair research Solutions GmbH	0	16
5,545	SphingoTec GmbH	0	14
5,547	Spectral Service AG	0	14
5,530	Max-Planck-Forschungsstelle fur die Wissenschaft der Pathogene	0	12
5,535	Bard College Berlin ? A Liberal Arts University	0	9
5,546	WSG - Westdeutsche Studiengruppe GmbH	0	9
5,508	EUROfusion Programme Management Unit Garching	0	8
1,504	EnBW Energie Baden-Wurttemberg AG	0	7
4,176	MEDA Pharma GmbH	0	7

Table 12: Institutions with no publications in 2020 in wos_b_2021 that had more than 5 publications in 2019 in the wos_b_2020 database.

PK_KB_INST	Name	Previous pubs	Current pubs
1,318	Oncotest GmbH - Institute for Experimental Oncology	6	0
1,355	metanomics GmbH	8	0

Table 13: Institutions with more than 20 publications in 2019 in the wos_b_2020 that increased in publication counts by over 40% to 2020 in the wos_b_2021 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
1,134	Fraunhofer-Institut fur Toxikologie und Experimentelle Medizin	45	228	183	406.7
4,737	Institute for Advanced Sustainability Studies e.V.	38	86	48	126.3
5,324	Mercator Research Institute on Global Commons and Climate Change	32	68	36	112.5

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
5,368	Translational Lung Research Center Heidelberg	36	72	36	100.0
179	ESCP Europe Wirtschaftshochschule Berlin e.V.	21	41	20	95.2
5,252	Medizinische Hochschule Brandenburg Theodor Fontane	70	126	56	80.0
177	Hertie School of Governance	46	82	36	78.3
4,198	Hasso-Plattner-Institut fur Softwaresystemtechnik (HPI)	35	62	27	77.1
5,478	Leibniz-Institut fur Werkstofforientierte Technologien - IWT	34	60	26	76.5
582	Hochschule Mittweida, University of Applied Sciences	21	37	16	76.2
198	Senckenberg Forschungsinstitut und Naturmuseum Frankfurt	208	357	149	71.6
625	Hochschule Furtwangen - Informatik, Technik, Wirtschaft, Medien, Gesundheit	46	78	32	69.6
1,115	Fraunhofer-Institut fur Holzforschung	23	39	16	69.6
713	Institut fur Herzinfarktforschung (IHF)	33	55	22	66.7
569	Hochschule Osnabruck	50	82	32	64.0
756	Forschungszentrum caesar	25	41	16	64.0
586	Hochschule Magdeburg-Stendal	32	52	20	62.5
1,603	CSL Behring GmbH	21	34	13	61.9
362	Klinikum St. Georg	31	49	18	58.1
506	Unfallkrankenhaus Berlin (UKB)	26	41	15	57.7
1,116	Fraunhofer-Institut fur Umwelt-, Sicherheits- und Energietechnik	42	66	24	57.1
211	Senckenberg Museum fur Naturkunde Gorlitz	29	45	16	55.2
743	Deutsches Rotes Kreuz e.V.	90	139	49	54.4
316	Klinikum Oldenburg gGmbH	24	37	13	54.2
5,348	Deutsches Zentrum fur Lungenforschung	287	441	154	53.7
602	Ernst-Abbe-Hochschule Jena ? University of Applied Sciences	29	44	15	51.7
634	Frankfurt University of Applied Sciences	34	51	17	50.0
5,293	Biomedical Research in Endstage and Obstructive Lung Disease (BREATH)	44	66	22	50.0
1,144	Fraunhofer-Institut fur Physikalische Messtechnik	22	32	10	45.5
2,121	Fraunhofer-Institut fur Mikrostruktur von Werkstoffen und Systemen	67	97	30	44.8
384	Stadtisches Klinikum Karlsruhe gGmbH	34	49	15	44.1

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
626	Hochschule Fulda - University of Applied Sciences	34	49	15	44.1
5,210	Berliner Institut fur Gesundheitsforschung	1,139	1,631	492	43.2
1,008	Max-Planck-Institut fur Wissenschaftsgeschichte	21	30	9	42.9
1,171	Fraunhofer-Institut fur Intelligente Analyse- und Informationssysteme	21	30	9	42.9
144	Zeppelin Universitat - Hochschule zwischen Wirtschaft, Kultur und Politik	24	34	10	41.7
39	Leibniz-Institut fur Kristallzuchtung (IKZ)	80	113	33	41.2
640	Hochschule fur nachhaltige Entwicklung Eberswalde	34	48	14	41.2

Table 14: Institutions with more than 20 publications in 2019 in the wos_b_2020 that decreased in publication counts by over 40% to 2020 in the wos_b_2021 database.

PK_KB_INST	Name	Previous pubs	Current pubs	No. diff.	Perc. diff.
493	Klinikum Links der Weser	22	13	-9	-40.9
48	Leibniz-Institut fur Atmospharenphysik e.V. an der Universitat Rostock (IAP)	38	22	-16	-42.1
736	European Space Operations Centre	30	17	-13	-43.3
1,541	Daimler AG	58	32	-26	-44.8
552	Technische Hochschule Wildau (FH)	38	20	-18	-47.4
4,411	Klinikum Frankfurt Hochst GmbH	29	15	-14	-48.3
646	Hochschule fur angewandte Wissenschaften Coburg	41	21	-20	-48.8
1,166	Fraunhofer-Institut fur Bauphysik	30	15	-15	-50.0
372	Klinikum Augsburg	76	23	-53	-69.7
5,203	Nanosystems Initiative Munich (NIM)	74	16	-58	-78.4

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_extended discipline in 2019 in both databases, and in the right panel the median number of authors per discipline in 2019 in the wos_b_2021 database compared to 2020 in the wos_b_2021 database.

While little change is expected to be seen in the left-hand panel of Figure 21 as the number of authors on a paper is unlikely to change between databases, differences in the right-hand panel indicate potential changes in disciplines' collaboration patterns. Disciplines for which the median number of authors changed by more than 1, based on the right-hand panel of Figure 21, are shown in Table 15. Also, to assess trends over a longer time-series and the full range of authors, we present the percentage of publications in each quartile of the range of authors in each Research Area over the most recent years of both databases in Figure 22.

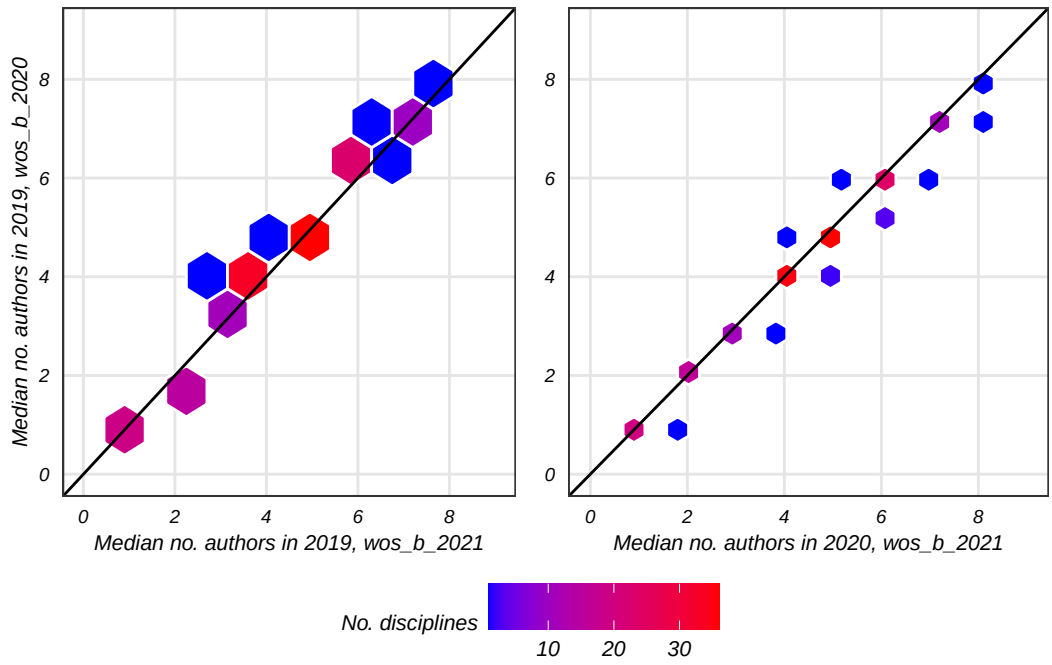

Figure 21: Median number of authors per discipline (sc_extended) between databases, where colour denotes the number of disciplines with this combination of mean authors.

Table 15: Disciplines (sc_extended) where the median number of authors changed by more than 1 between the last common year in the previous database and the latest year in the current database.

Discipline	Previous median authors	Current median authors	Diff.

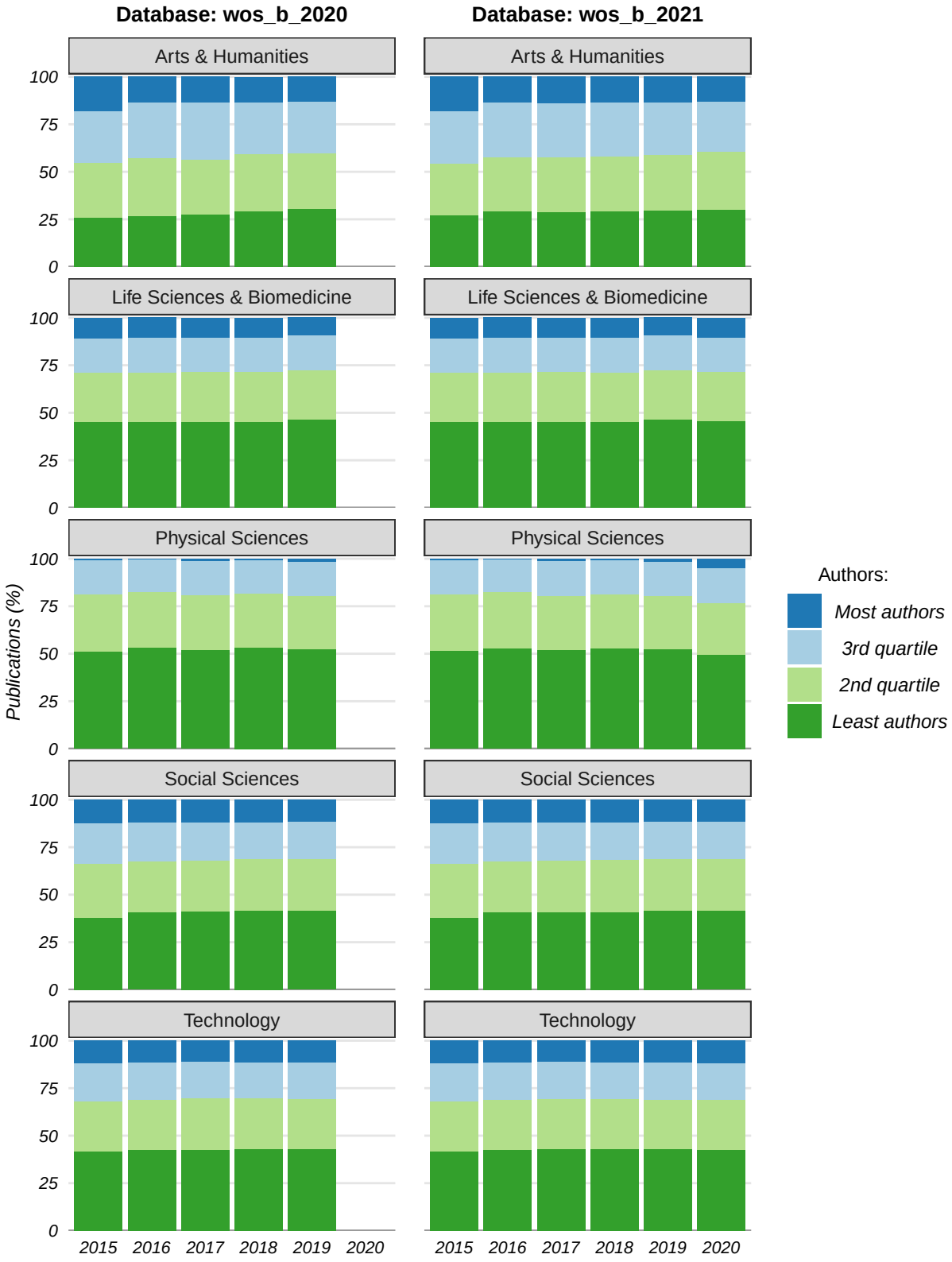

Figure 22: Distribution of publications over quartiles based on number of authors over time.

Source items: Percentage by 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 not indexed and therefore non-source. Only source items are included in citation counts and so understanding the percentage of items cited that are also source can give an indication of the depth of WoS' coverage of a discipline. That is, if a large number of indexed items' sources are not indexed, the reverse is also likely true and a large number of citations of indexed items are also missing, which has the effect of reducing citation counts for disciplines with lower coverage, such as the arts and humanities.

The percentage of references that are source items is expected to increase over time as Clarivate Analytics 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 23 the percentage of references that are source items per sc_extended discipline in 2019 in both databases, and in the right-hand panel the percentage of references that are source items per discipline in 2019 in the wos_b_2021 database compared to 2020 in the wos_b_2021 database.

It is in the right-hand panel that the effect of recently indexed journals may become apparent, where an increase in the percentage of source items may be seen if the journal is often cited within a discipline. The disciplines with a change in the percentage of indexed references of more than five percentage points between databases, based on the right-hand panel of Figure 23, are shown in Table 16. Longer term trends can be seen in Figure 24 where we present the percentage of reference that are source items per Research Area over the last ten common years of both databases.

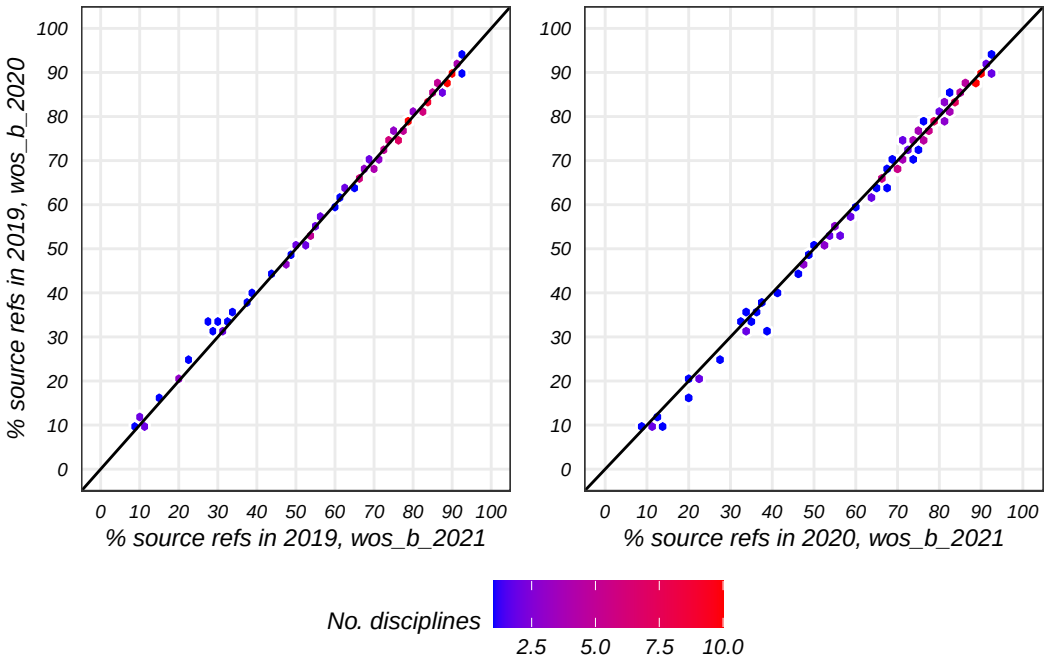

Figure 23: The percentage of cited items that are source items per sc_extended discipline by database, where colour denotes the number of disciplines with this combination of source references.

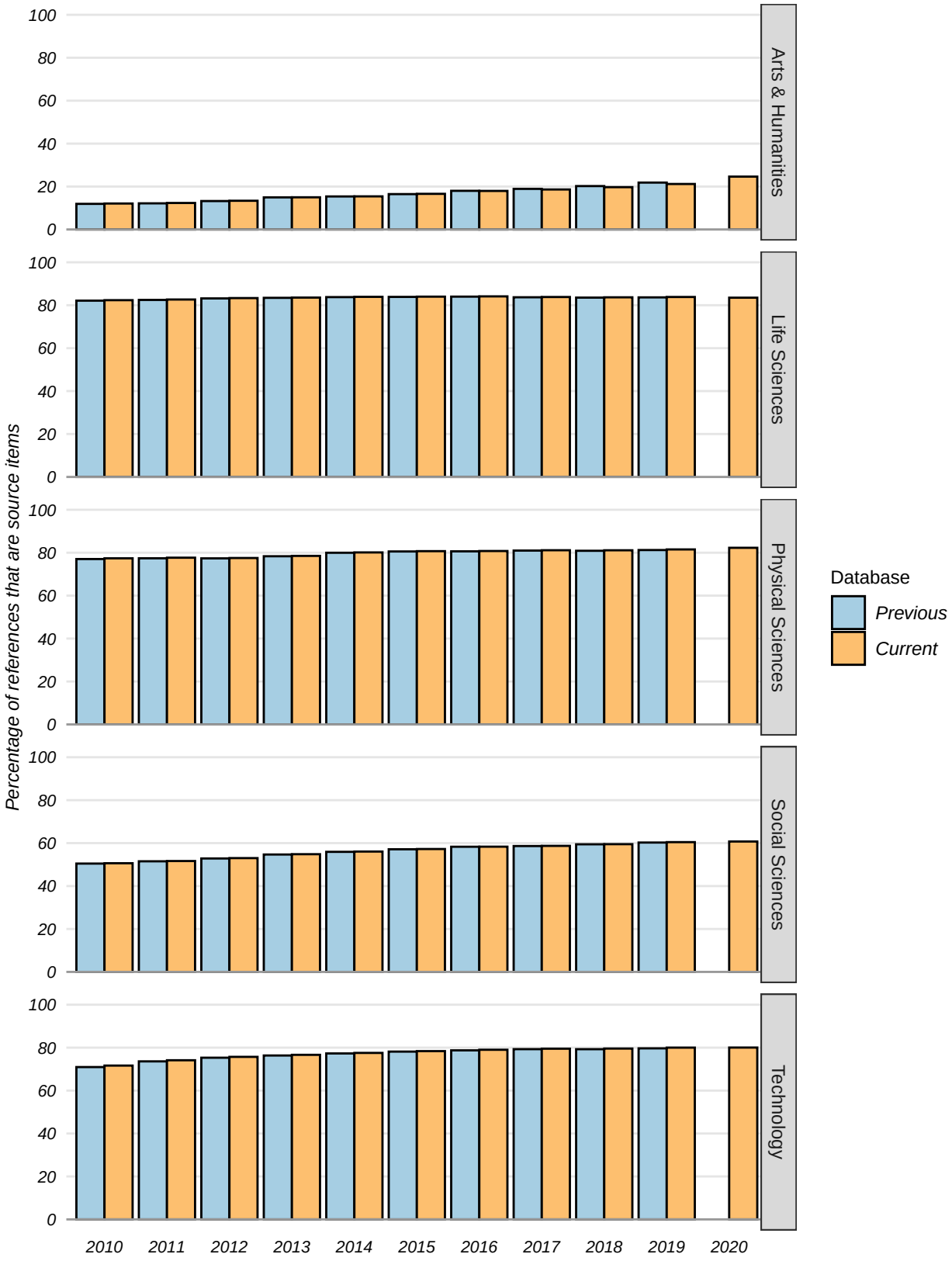

Figure 24: The percentage of references that are source items by Research Area and database over time.

Table 16: Disciplines (sc_extended) where the percentage of indexed references changed by 3 or more percentage points between 2019 in wos_b_2020 and 2020 in wos_b_2021.

Discipline	Previous no. refs.	Current no. refs.	Prvs % source	Crrnt % source	Change
Music	223,320	290,782	31.1	38.2	7.1
Religion	665,548	725,908	24.8	28.7	3.9
Dance	7,095	13,973	16.3	20.1	3.8
Philosophy	732,665	1,021,740	30.5	34.0	3.5
Mycology	2,008,133	2,281,378	75.6	71.7	-3.9

References

[1] S. Stahlschmidt, D. Stephen and S. Hinze. “Performance and Structures of the German Science System”. In: Studien zum deutschen Innovationssystem. Expertenkommission Forschung und Innovation (EFI), 2019. Chap. Studie 5-2019.

[2] J. Wang. “Citation time window choice for research impact evaluation”. In: *Scientometrics* 94.3 (2013). doi:10.1007/s11192-012-0775-9, pp. 851–872.