Which of the metadata with relevance for bibliometrics are the same and which are different when switching from *Microsoft Academic Graph* to *OpenAlex*?

Thomas Scheidsteger; Robin Haunschild

Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87295

Recommended citation:

Scheidsteger, Thomas; **Haunschild, Robin** (2023). "Which of the metadata with relevance for bibliometrics are the same and which are different when switching from *Microsoft Academic Graph* to *OpenAlex*?". *Profesional de la información*, v. 32, n. 2, e320209.

https://doi.org/10.3145/epi.2023.mar.09

Article received on January 18th 2023
Approved on February 8th 2023

Thomas Scheidsteger ✉
https://orcid.org/0000-0001-8351-2498

Max Planck Institute for Solid State Research
IVS-CPT
Heisenbergstr. 1
70569 Stuttgart, Germany
T.Scheidsteger@fkf.mpg.de

Robin Haunschild
https://orcid.org/0000-0001-7025-7256

Max Planck Institute for Solid State Research
IVS-CPT
Heisenbergstr. 1
70569 Stuttgart, Germany
R.Haunschild@fkf.mpg.de

Abstract

With the announcement of the retirement of *Microsoft Academic Graph* (*MAG*), the non-profit organization *OurResearch* announced that they would provide a similar resource under the name *OpenAlex*. Thus, we compare the metadata with relevance to bibliometric analyses of the latest *MAG* snapshot with an early *OpenAlex* snapshot. Practically all works from *MAG* were transferred to *OpenAlex* preserving their bibliographic data publication year, volume, first and last page, DOI as well as the number of references that are important ingredients of citation analysis. More than 90% of the *MAG* documents have equivalent document types in *OpenAlex*. Of the remaining ones, especially reclassifications to the *OpenAlex* document types journal-article and book-chapter seem to be correct and amount to more than 7%, so that the document type specifications have improved significantly from *MAG* to *OpenAlex*. As another item of bibliometric relevant metadata, we looked at the paper-based subject classification in *MAG* and in *OpenAlex*. We found significantly more documents with a subject classification assignment in *OpenAlex* than in *MAG*. On the first and second level, the classification structure is nearly identical. We present data on the subject reclassifications on both levels in tabular and graphical form. The assessment of the consequences of the abundant subject reclassifications on field-normalized bibliometric evaluations is not in the scope of the present paper. Apart from this open question, *OpenAlex* seems to be overall at least as suited for bibliometric analyses as *MAG* for publication years before 2021 or maybe even better because of the broader coverage of document type assignments.

Keywords

Subject classification; Fields of study; Concepts; Bibliographic data; Metadata; Document types; Citation analysis; *Microsoft Academic Graph*; *MAG*; *OpenAlex*; Bibliometrics.

Acknowledgements

We thank Jason Priem for very helpful comments on an earlier draft. The present study is based on the conference contribution (Scheidsteger; Haunschild, 2022) to the *STI 2022* congress (Granada, Spain), but especially augmented by a much more detailed investigation of the subject classifications in both databases.

1. Introduction

Since its launch in 2015, *Microsoft Academic Graph* (*MAG*; **Sinha** *et al.*, 2015) had been a promising new data source for bibliometric analyses due to its large coverage and set of available metadata (**Harzing**; **Alakangas**, 2017). Therefore, *MAG* has been the object of many studies, in particular comparisons with other important bibliographic databases. In one of the last and thus far largest ones, **Visser**, **Van-Eck**, and **Waltman** (2021) compared *MAG* with *Web of Science*, *Scopus*, *Dimensions*, and *Crossref*.

In May 2021, it was announced by the *Microsoft Blog* (2021) that the *Microsoft Academic* website, application programming interfaces (API), and snapshots would retire on December 31, 2021. Soon after that, the non-profit organization *OurResearch*, aiming at providing "a fully open catalog of the global research system" (*OurResearch*, 2021), announced they would preserve and incorporate the last full *MAG* corpus, only excluding patent data, and to continue and hopefully improve it. Another main source of data should be *Crossref*. In January 2022, *OpenAlex* (*http://openalex.org*) was launched and provided API access to their services as well as data dumps for any purposes. The *Curtin University's Open Knowledge Initiative* (*COKI*) has already started to monitor the development of *OpenAlex*, in particular assessing and comparing the value added by *OpenAlex* to *MAG* and to *Crossref*, both in coverage of publications and other research output (**Kramer**, 2022).

Scheidsteger *et al.* (2018) studied the possibility of using *MAG* data for the calculation of field- and time-normalized citation scores. They compared the scores derived from subject classifications and coverage in *MAG* to those derived from subject categories and coverage in *Web of Science* (*WoS*). In the present study, we are interested in comparing metadata that are relevant for bibliometric analyses (in particular field and time normalization of citations) of *MAG* and *OpenAlex*:

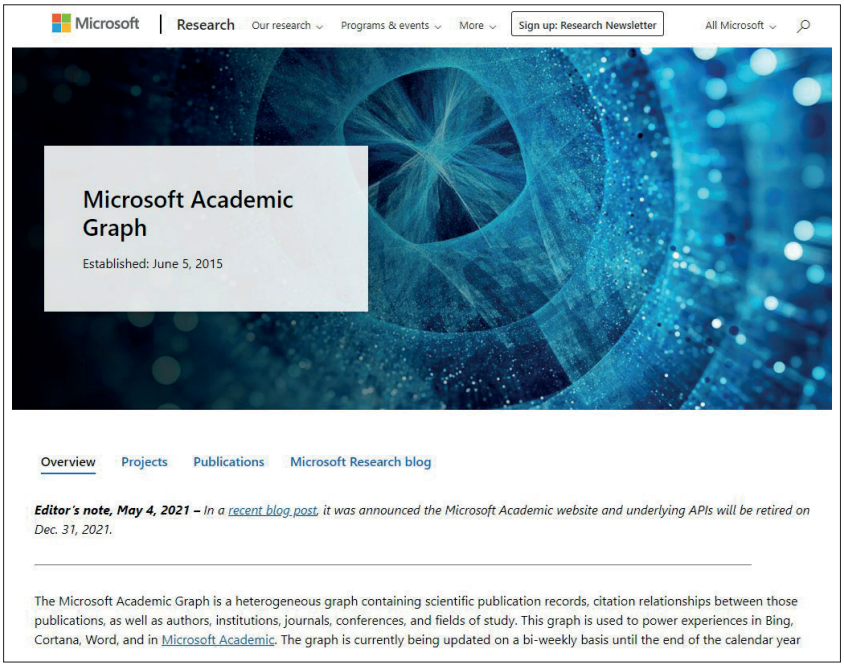
- the coverage of documents over the years,
- the agreement of bibliographic data,
- the numbers of references of each document,
- the kind and distribution of document types,
- the distribution of and relation between subject classifications.

2. Data and methods

Microsoft Academic Graph (*MAG*)

We downloaded the *Microsoft Academic Graph* (*MAG*) dataset via the *Microsoft Azure* portal at the end of December 2021 and received data timestamped with 6 December 2021 (**Sinha** *et al.*, 2015). See: *https://www.microsoft.com/en-us/research/project/academic*

We were not able to get newer data at the beginning of 2022 after the official expiration date of the *MAG* service. According to the *OpenAlex migration guide* (*OpenAlex*, 2021), no patents have been transferred from *MAG* to *OpenAlex*. Therefore, we excluded all items with document type *Patent* from the comparison. In order to facilitate the distinction between the two databases, we keep the case of the document type names as they are used in both databases. In particular, *MAG* types are capitalized. Because *MAG* data do not contain the full year 2021, we restricted our analyses to the publication years before 2021. Thus, we considered 197,445,041 papers in *MAG* of which 95,160,734 possess a DOI.



https://www.microsoft.com/en-us/research/project/microsoft-academic-graph

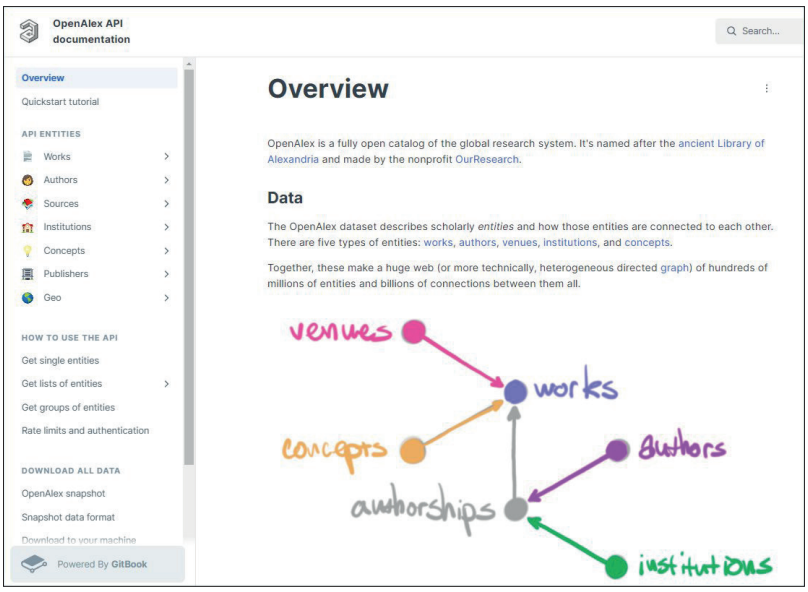
OpenAlex

The *OpenAlex* data dump was retrieved on 9 February 2022 with an update timestamp of 31 January 2022 on the main table (*works*). Both datasets were imported into and processed in our locally maintained *PostgreSQL* database at the *Max Planck Institute for Solid State Research* (Stuttgart, Germany). Before the publication year 2021, we have a total of 198,606,165 works in *OpenAlex*, of which 96,268,256 possess a DOI.

Documents in *MAG* and *OpenAlex* can be linked via a unique ID. *OpenAlex* like *MAG* only contains linked references. For most works, there are "Fields of Study" available –called "concepts" in *OpenAlex* and (only there) all linked to a respective *Wikidata* ID via the table *concepts*. For more details on the approach and the structure of *OpenAlex* see **Priem**, **Piwowar**, and **Orr** (2022).

Software

The statistical evaluations have been done by using *R* (*R Core Team*, 2020), the graphical presentation in Figure 1 by using the *R* package *ggplot2* (**Wickham**, 2016), and the alluvial plots Figure 2, Figure 3, Figure 4, and Figure 5 by using the *R package alluvial* (**Bojanowski**; **Edwards**, 2016).



https://docs.openalex.org

3. Results

Coverage of publication years

Only 777 IDs from *MAG* are not incorporated in *OpenAlex*, starting with one item in 1952 and reaching a maximum of 201 in 2020. The document types in *MAG* of these missing items are about 40% *Journal* and *None*, each, and about 15% *BookChapter.* Over the whole period since 1952, of the 777 *MAG* IDs, 654 have DOIs, only two of them could not be found in *Crossref*. 347 of these DOIs contain the ISBN Bookland prefix "978" or "979" and therefore point to books or book chapters, but only one third of them is assigned to the types *Book* or *BookChapter* in *MAG*. As expected, the number 777 of missing *MAG* IDs exactly matches the difference between the overall number of *MAG* papers and 197,444,264 *OpenAlex* works that have a *MAG* ID associated with them. Of the 654 DOIs, 23 had been associated with more than one *MAG* ID (two of them with three, the other ones with two) and –apart from one (*10.1016/j.physrep.2013.03.005*)– all could be found in *OpenAlex* and each had one occurrence less. In 16 cases the resp. preprint entry had been dropped in favor of the resp. journal article entry, and in two cases the resp. entry as journal article had been dropped in favor of the resp. entry as book chapter.

There are 1,161,901 works indexed in *OpenAlex* that have *no* corresponding record in *MAG*, 1,108,176 of them having a DOI in *OpenAlex*, in particular 1,877 documents before 1800, the first publication year in *MAG*. In the following, only the documents both databases have in common are going to be investigated.

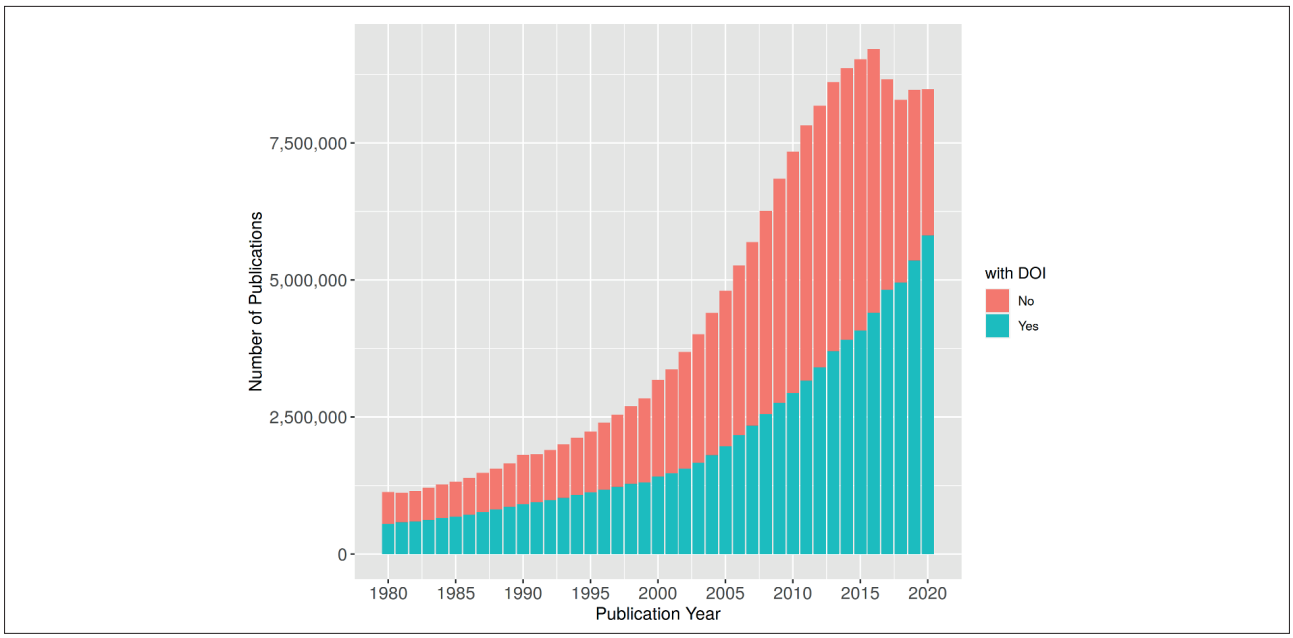


Figure 1. Numbers of common *OpenAlex-MAG* documents across the years 1980 to 2020

Figure 1 shows the annual numbers of common documents with and without DOI across the years 1980 until 2020. The unexpected decrease of the total number starting in 2017 is due to the shrinking number of documents without a DOI which in turn is by far dominated by the number of documents with *no* document type assigned.

> Unexpectedly, the total number of documents starts to decrease in 2017. This is due to a shrinking number of documents without a DOI which in turn is by far dominated by the number of documents with *no* document type assigned

Comparison of bibliographic data

For the 197,444,264 documents in *OpenAlex* with an ID in *MAG*, we firstly check if the bibliographic data from *MAG*, like volume, issue, first page, last page, and DOI are preserved after the transfer to *OpenAlex*. When volume or issue were available in *MAG* these data have been completely transferred to *OpenAlex*. This is also the case for first and last pages and DOIs. During our investigation, we found some issues with the data quality:

(i) In more than 28,800 cases, the fields "first page" and "last page" contained not a single number but the same range of numbers, e.g., "35-46".

(ii) More than 810,028 DOIs occur *more than once* in the dataset, 7,626 of them at least ten times, and 235 at least 100 times. Of the top 100 most-frequently occurring DOIs, only 29 can be resolved.

(iii) More than 6,000 DOIs contain non-latin characters, less than 200 could be resolved.

Secondly, concerning the number of (linked) references for a document, we counted the entries in *OpenAlex*'s table of references for each work (*works_referenced_works)* and found *no* deviation from the respective values (*nref*) in the corresponding table of *MAG*. But *nref* had been calculated including patent references. Obviously, in *OpenAlex*, the references to patents had been kept but not the patent documents themselves.

Document types

In *MAG*, we are dealing with seven document types: *Book*, *BookChapter*, *Conference*, *Dataset*, *Journal*, *Repository*, and *Thesis*. Table 1 lists the numbers and shares of their occurrences. Nearly 45% of the documents are classified as *Journal*, but nearly the same number of documents have *no* document type assigned (*None*).

In *OpenAlex*, there are 26 document types that inherit their definition from another major data source *Crossref* –as documented in *Crossref*'s *content type markup guide* (*Crossref*, 2021). Obviously, all works in *OpenAlex* with a *Crossref* DOI receive their document type from there. Those document types with a share of more than 0.1% of all documents are listed in Table 2. There are additional nine million items in *OpenAlex* assigned to the document type *journal-article* as compared to the *MAG* document type *Journal*. The *OpenAlex* items of document type *journal-article* cover nearly one half of all documents, but the items without a document type (*none*) are still more than a third of all. However, the document types *Journal* and *None* occur about equally frequently in *MAG*. The increased numbers of journal articles, conference proceedings, and book chapters are especially interesting from a bibliometric point of view.

As displayed in Table 3, about 90.1% of all items have the obviously equivalent document types in both databases.

Table 1. Number and percentages of document types in *MAG*

Document types in *MAG*	Number of items	Percentage of items
Journal	87,430,385	44.28
None	85,844,335	43.48
Thesis	5,925,439	3.00
Conference	5,053,232	2.56
Repository	4,779,269	2.42
Book	4,588,285	2.32
BookChapter	3,691,552	1.87
Dataset	132,544	0.07
Sum	197,445,041	100.00

Table 2. Numbers and percentages of document types in *OpenAlex*

Document types in *OpenAlex*	Number of items	Percentage of items
journal-article	96,547,138	48.61
none	70,155,602	35.32
book-chapter	9,588,895	4.83
proceedings-article	7,051,207	3.55
dissertation	6,126,640	3.08
book	4,522,989	2.28
posted-content	3,093,874	1.56
report	464,164	0.23
dataset	276,311	0.14
monograph	212,401	0.11
other types	566,944	0.29
sum	198,606,165	100.00

Table 3. Shares of transfers of equivalent document types between *MAG* and *OpenAlex*

MAG	*OpenAlex*	Number of items	Percentage of items
Journal	*journal-article*	86,395,430	43.76
None	*none*	70,154,418	35.53
Thesis	*dissertation*	5,917,802	3.00
Book	*book*	4,421,867	2.24
Conference	*proceedings-article*	4,285,360	2.17
BookChapter	*book-chapter*	3,662,705	1.86
Repository	*posted-content*	3,018,186	1.53
Dataset	*dataset*	132,421	0.07
Sum		177,988,189	90.15

The more interesting cases are the reclassifications. Therefore, we show in Figure 2 an alluvial diagram of the corresponding document types in both databases, *excluding* the transfers from Table 3.

Those reclassifications occurring in relevant numbers that sum up to nearly 9.3% of all documents, are listed in Table 4. In order to get an impression of the quality of these reclassifications, we add some characteristics of respective random samples of ten documents, each. All of them had a DOI –as we could expect because of *Crossref* being the main source of document type information. Indeed, less than 10,000 documents *without* a DOI have been reclassified, i.e., about 0.05% of all 20 million reclassifications.

> The reclassification to type *book-chapter* in *OpenAlex* seems to work fairly well. This is also the case for *journal-article*. In particular, many documents using non-latin character sets are now getting classified, and a substantial number of items with DOIs and the document type *Repository* in *Microsoft Academic Graph* are correctly recognized as *journal-article*

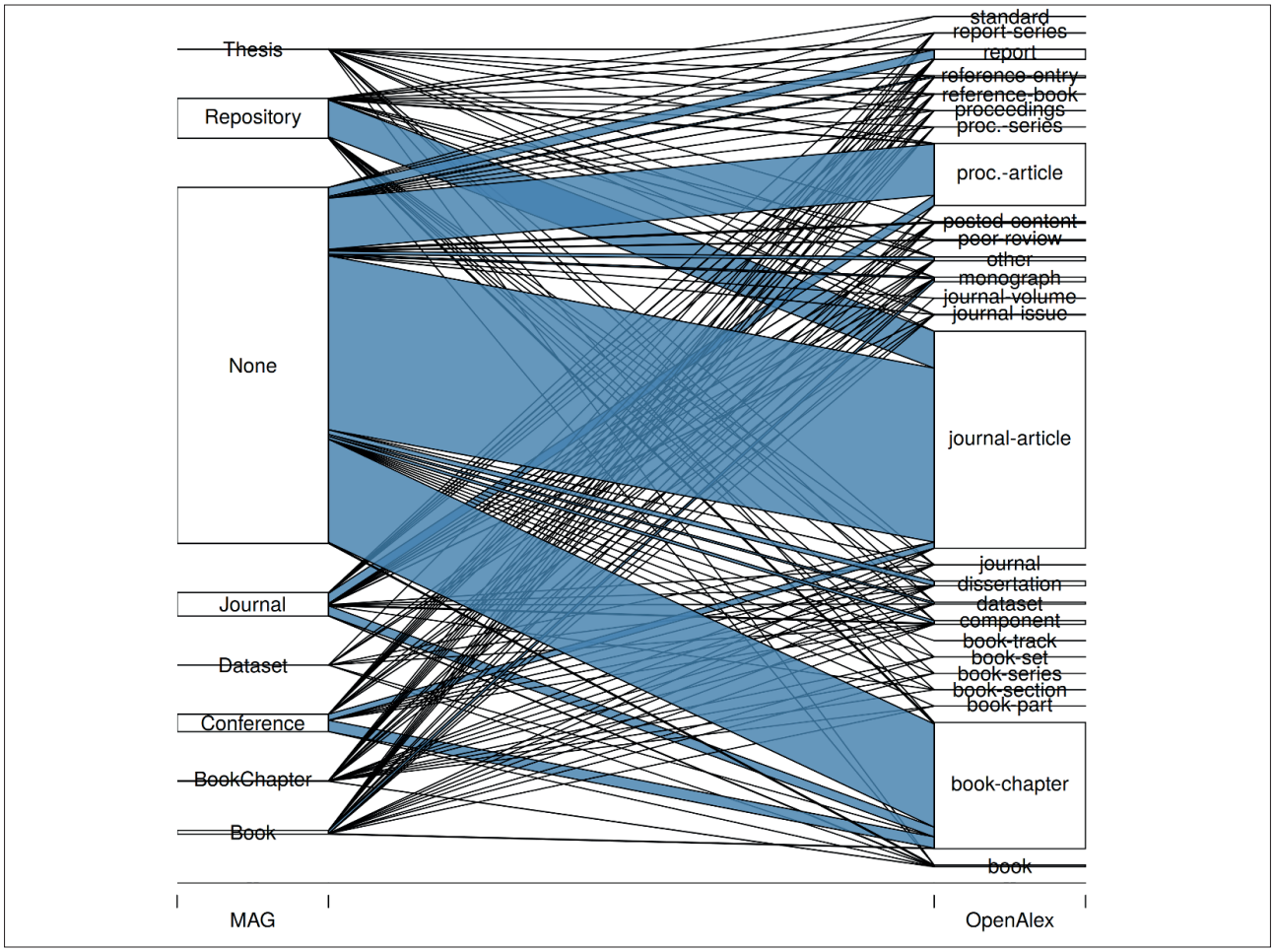


Figure 2. Alluvial diagram of document type reclassifications from *MAG* to *OpenAlex*

Table 4. Shares of reclassifications of document types from *MAG* to *OpenAlex* together with some characteristics of corresponding random samples of ten documents. Shares of at least 0.1% are shown.

Document types			Random samples of ten documents	
MAG	***OpenAlex***	**Percentage of all documents**	**Span of publication years**	**Other characteristics**
None	*journal-article*	3.88%	1928 - 2018	Eight titles with Cyrillic, far-eastern, or Arab character set; one Dutch document with English title;
None	*book-chapter*	2.30%	1984 - 2020	All DOIs containing the Bookland prefix "978"; one German title
None	*proceedings-article*	1.14%	1971 - 2019	Seven Cyrillic or Arab titles; only two conference papers identifiable
Repository	*journal-article*	0.82%	1988 - 2016	Four *ChemInform Abstracts*; five *arXiv* papers: All DOIs point to published papers; one *SSRN* preprint from 2012, published in 2016 in a journal
Conference	*book-chapter*	0.25%	2001 - 2020	All published in conference proceedings by *Springer* as part of a book series; eight DOIs contain Bookland prefix "978"; seven documents from LNCS; only one document noted by *Springer* as chapter, the others as conference papers
Journal	*proceedings-article*	0.23%	2010 - 2017	Three poster presentation abstracts in the supplement of a journal; four documents from the *Proceedings of SPIE*; two documents in proceedings of a medical conference as supplement to a journal.
Journal	*book-chapter*	0.22%	1965 - 2017	All book chapters; eight DOIs contain Bookland prefix "978"
Conference	*journal-article*	0.14%	1987 - 2014	No conference papers; four publishers incorrect in *MAG*
None	*report*	0.20%	1964 - 2019	Seven technical reports or geological survey data from US and Canadian governments
None	*dissertation*	0.10%	1973 - 2018	Theses and dissertations at institutional repositories (five US, four Brasilian, one Greek)

The reclassification to type *book-chapter* in *OpenAlex* seems to work fairly well. This is also the case for *journal-article*. In particular, many documents using non-latin character sets are now getting classified, and a substantial number of items with DOIs that *MAG* had labelled as *arXiv* preprints with document type *Repository* are correctly recognized as *journal-article*. On the other hand, the assignment of *ChemInform* abstracts to this document type is debatable, but they are definitely no preprints. Conference papers seem to be a special case: Documents incorrectly assigned to *Journal* get corrected to *proceedings-article*, but for documents without a document type in *MAG* the assignment of *proceedings-article* is not that accurate or at least difficult to verify. In case of *MAG* type *Conference*, the reclassification to *journal-article* seems to be overall correct, whereas the reclassification of *Lecture notes in computer science* (*LNCS*) contributions to *book-chapter* seems to be the result of their appearance as part of book series and of the format of their DOIs containing the Bookland prefix "978" (*DOI.org*, 2019). This fact should be kept in mind for bibliometric studies in computer sciences, which probably should include book chapters as well.

Subject classifications

OpenAlex states in their migration guide (*OpenAlex*, 2021) that they use the same taxonomy as *MAG* but have reduced the number of "Fields of Study" (FoS) by removing those with less than 500 papers associated. Moreover, they have applied a different algorithm, i.e., model *V1*, that used paper titles and a few other features, but not abstract data. The latter were only used later in 2022 with the implementation of model *V2* of their open-source software (**Priem**; **Piwowar**, 2022).

A quick look reveals the persistence of all 19 top-level FoSs (level=0) from *MAG* as well as of 284 of the 292 FoSs of the next level (level=1). Table 5 lists the distribution of all FoS levels from 0 to 5 in both databases. The strongest reduction of FoS numbers occurs in the levels 3 to 5 where less than 10% persist. However, in both databases, *MAG* and *OpenAlex*, the granularity does not necessarily increase with the FoS level. Level 3 has the highest number of FoSs. The total number of FoSs on all levels is 714,971 in *MAG* and only 65,073 in *OpenAlex*, which means a reduction to 9.1%. Interestingly, in levels 2 to 5, a substantial number of FoSs have less than 500 works assigned to them, e.g., more than 4,000 FoSs on levels 2 and 3, respectively.

Table 5. Distribution of FoSs in *MAG* and *OpenAlex*

Level	#*MAG*	#*OpenAlex*	Difference (#*MAG* - #*OpenAlex*)	Percentage (#*OpenAlex*/#*MAG**100)
0	19	19	0	100.00
1	292	284	8	97.26
2	137,415	21,460	115,955	15.62
3	330,275	24,768	305,507	7.50
4	134,843	12,406	122,437	9.20
5	112,127	6,136	105,991	5.47
All levels	714,971	65,073	649,898	9.10

Top-level Fields of Study

Even if the top-level FoSs persist, they are very differently associated to the papers. For example, one paper (accessed on 26 April 2022) had one top-level FoS and one level-1 FoS in *MAG*, but it has six additional top-level FoSs and one additional level-1 FoS in *OpenAlex*:
https://api.openalex.org/works/W2178938397

Table 6 shows some statistical measures of the common publication set concerning the number of papers with a FoS assigned. The total number of papers with any FoS is significantly increased: 30.5 of 48.9 million documents without any FoS in *MAG* have at least one FoS in *OpenAlex* –28.8 million having a top-level FoS in *OpenAlex*– so that the coverage increases from 74.6% to 86.6%. The number of assignments per paper to a top-level FoS is drastically increased in *OpenAlex*: About 147 million papers in *MAG* and about 171 million papers in *OpenAlex* have at least one top-level FoS assigned to them. Of those papers, 65 thousand in *MAG* and 53 million in *OpenAlex* have multiple top-level FoSs (up to seven) assigned to them. Thus, by applying the concept algorithm of *OpenAlex*, the multiple assignment to top-level FoSs proliferated.

Table 6. Numbers and percentages of documents with FoS assigned

Statistical measures of common publication set	*MAG*	*OpenAlex*
Number of documents	197,444,264	197,444,264
Number of assignments to any FoS	1,092,748,572	1,095,801,888
Number of documents with any FoS	148,518,539	175,993,558
Number of documents without any FoS	48,925,725	21,450,706
Coverage of documents with any FoS	75.22%	89.14%
Mean assignments to any FoS per document	7.36	6.23
Number of assignments to a top-level FoS	147,426,219	229,560,450
Number of documents with a top-level FoS	147,360,860	170,900,225
Coverage of documents with a top-level FoS	74.63%	86.56%
Number of documents with multiple top-level FoS assignments	65,359	52,966,153
Percentage of documents with multiple top-level FoS assignments of all documents with FoS assignments	0.04%	30.99%
Mean assignments to a top-level FoS per document	1.000444	1.343243

Table 7 compares the number of assignments to top-level FoSs in both databases. The highest relative increase is to be seen with Arts and the strongest decrease with Engineering.

Table 7. Comparison of the numbers of top-level FoS assignments in *MAG* and *OpenAlex*

FoS	# *MAG*	% *MAG*	# *OpenAlex*	% *OpenAlex*	% *OpenAlex* / % *MAG*
Art	3,717,975	2.52	12,873,508	5.61	2.22
Biology	13,169,649	8.93	14,242,938	6.20	0.69
Business	5,174,422	3.51	12,010,241	5.23	1.49
Chemistry	14,191,693	9.63	20,029,716	8.73	0.91
Computer science	12,312,525	8.35	25,678,965	11.19	1.34
Economics	3,130,346	2.12	4,064,798	1.77	0.83
Engineering	8,472,749	5.75	3,117,013	1.36	0.24
Environmental science	3,533,640	2.40	6,702,031	2.92	1.22
Geography	4,447,923	3.02	7,053,608	3.07	1.02
Geology	3,061,102	2.08	3,621,249	1.58	0.76
History	3,059,007	2.07	4,454,112	1.94	0.94
Materials science	11,063,791	7.50	17,437,416	7.60	1.01
Mathematics	6,021,856	4.08	6,101,501	2.66	0.65
Medicine	27,897,600	18.92	36,085,634	15.72	0.83
Philosophy	2,010,846	1.36	5,916,152	2.58	1.89
Physics	6,873,294	4.66	9,938,209	4.33	0.93
Political science	6,775,718	4.60	17,760,011	7.74	1.68
Psychology	8,063,945	5.47	13,966,595	6.08	1.11
Sociology	4,448,138	3.02	8,506,753	3.71	1.23
All assignments	147,426,219		229,560,450		

Of the nearly 49 million documents without any FoS in *MAG*, nearly 29 million received at least one top-level FoS in *OpenAlex*. Table 8 shows the distribution of document types in *MAG* across both sets. About 80% of the papers without any FoS have no document type and half of them receive a top-level FoS in *OpenAlex* amounting to two thirds of all these documents. Considering the bibliometrically most interesting document types *Journal*, *BookChapter*, *Book*, and *Conference*, about 90% have a top-level FoS in *OpenAlex* amounting to a quarter of all those documents.

Table 8. Distribution of document types in MAG across the documents without any FoS in *MAG*, and share of documents with some top-level FoS in *OpenAlex*

Document type in *MAG*	#*MAG* no FoS	#*MAG* no FoS, but top-level FoS in *OpenAlex*	Percentage of papers with a top-level FoS in *OpenAlex* but no FoS in *MAG*
None	38,273,234	19,266,767	50.34
Journal	5,102,461	4,566,743	89.50
Thesis	2,453,810	2,249,289	91.67
BookChapter	1,416,913	1,235,266	87.18
Book	1,356,096	1,241,835	91.57
Repository	251,384	228,099	90.74
Conference	58,961	54,332	92.15
Dataset	12,866	9,561	74.31
Sum resp. average	48,925,725	28,851,892	58.97

About 77.2% of all top-level assignments in *MAG* persist in *OpenAlex*, but this proportion varies significantly across the 19 top-level FoSs as Table 9 clearly shows –from less than a quarter for Engineering to more than 90% for Material Sciences and Medicine.

Table 9. Distribution of top-level FoSs in *MAG* and number and percentage of top-level FoSs persistent in *OpenAlex*.

FoS	#*MAG*	#*OpenAlex*	% persistent
Art	3,717,975	2,620,365	70.48
Biology	13,169,649	10,411,044	79.05
Business	5,174,422	4,200,803	81.18
Chemistry	14,191,693	12,194,451	85.93
Computer science	12,312,525	10,878,013	88.35
Economics	3,130,346	2,131,877	68.10
Engineering	8,472,749	2,023,815	23.89
Environmental science	3,533,640	2,712,884	76.77
Geography	4,447,923	2,366,289	53.20
Geology	3,061,102	2,302,537	75.22
History	3,059,007	1,650,999	53.97
Materials science	11,063,791	10,010,937	90.48
Mathematics	6,021,856	4,028,415	66.90
Medicine	27,897,600	25,953,084	93.03
Philosophy	2,010,846	1,240,834	61.71
Physics	6,873,294	5,517,376	80.27
Political science	6,775,718	4,899,049	72.30
Psychology	8,063,945	6,198,019	76.86
Sociology	4,448,138	2,520,233	56.66
All assignments	147,426,219	113,861,024	77.23

Figure 3 shows an alluvial plot of the transfer of paper-based subject classifications *without* the persistent FoS assignments of Table 9 so that the remaining reclassifications become more visible. Given the fact that *all* 342 possible reclassifications do indeed occur in our publication set, only the 94 connections with at least 200,000 occurrences are shown. Several reclassifications occur in comparable measures in both directions, e.g., in the pairs Sociology & Psychology, Sociology & Political Science, or Psychology & Medicine. Other ones show a significant transfer in mainly one direction, like from Engineering to Computer Science, from Mathematics to Computer Science, from Biology to Chemistry, or from Chemistry to Materials Science.

As an example, the distribution of assignments of the documents with top-level FoS Engineering in *MAG* to top-level FoSs in *OpenAlex* is given in Table 10. The right two columns indicate the share of the assignments to the respective FoS alone.

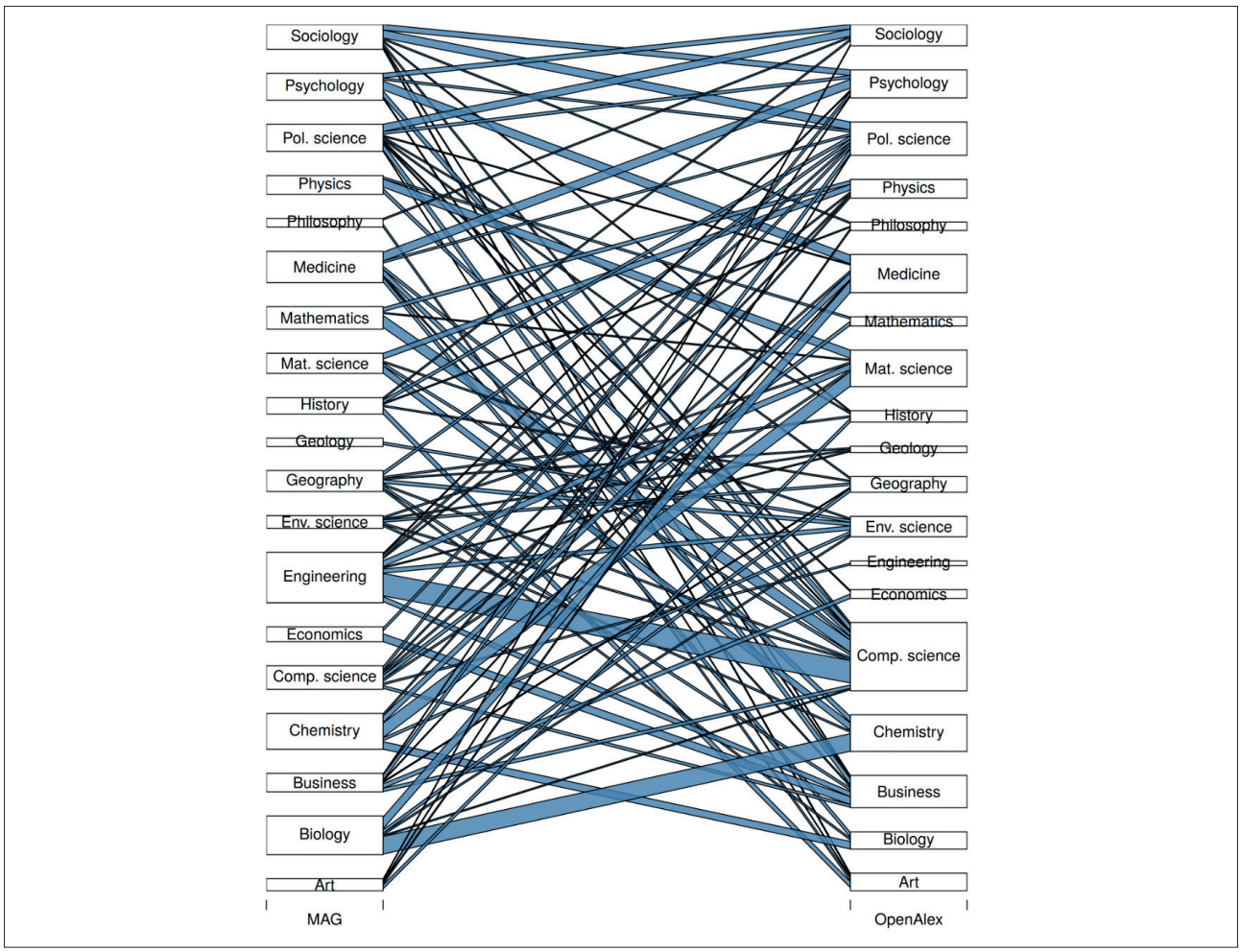


Figure 3. Alluvial diagram for the top-level FoS reclassifications from *MAG* to *OpenAlex*, showing only reclassifications that occur at least 200,000 times

Table 10. Distribution of assignments of the 8,472,749 documents with top-level FoS Engineering in *MAG* and their top-level FoSs in *OpenAlex*

FoS in *OpenAlex*	Number	Percentage	Number of papers that are assigned only to this FoS	Percentage of papers that are assigned only to this FoS
Art	155,994	1.84	83,922	53.80
Biology	37,572	0.44	21,505	57.24
Business	890,319	10.51	371,279	41.70
Chemistry	117,583	1.39	47,157	40.11
Computer science	3,880,097	45.80	2,199,873	56.70
Economics	52,304	0.62	10,490	20.06
Engineering	2,023,815	23.89	649,984	32.12
Environmental science	715,720	8.45	277,495	38.77
Geography	121,729	1.44	46,386	38.11
Geology	222,817	2.63	98,828	44.35
History	66,291	0.78	16,244	24.50
Materials science	1,028,047	12.13	524,787	51.05
Mathematics	122,422	1.44	20,450	16.70
Medicine	175,371	2.07	86,643	49.41
Philosophy	22,304	0.26	7,247	32.49
Physics	310,876	3.67	90,731	29.19
Political science	412,799	4.87	155,831	37.75
Psychology	245,947	2.90	99,949	40.64
Sociology	166,826	1.97	44,179	26.48
Sum/Average	10,768,833	127.10	4,852,980	45.07

Only nearly one quarter of the 8,472,749 documents that are assigned to Engineering in *MAG* is assigned to the same FoS in *OpenAlex* but nearly one half of them is assigned to the FoS Computer Science. There is a significant amount of multiple assignments in *OpenAlex* so that the total number of FoS assignments is increased by more than a quarter to 10,768,833.

Table 11 shows the top 10 most occurring lists of FoS assignments of Engineering papers from *MAG*. Individual FoSs are separated by a semicolon. Interestingly, the combination Computer science; Engineering comes second and, together with Engineering at the third place, it amounts to more than 75% of the assignments to Engineering in *OpenAlex*.

Table 11. Top 10 most occurring lists of top-level FoS assignments in *OpenAlex* for the 8,472,749 documents assigned to Engineering in *MAG*

Most frequent top-level FoS assignments in *OpenAlex*	**Number of papers with most frequent top-level FoS assignments in *OpenAlex***	**Percentage of papers with most frequent top-level FoS assignments in *OpenAlex***
Computer science	2,199,873	25.16
Computer science; Engineering	827,861	9.47
Engineering	649,984	7.43
Materials science	524,787	6.00
Business	371,279	4.25
Environmental science	277,495	3.17
Political science	155,831	1.78
Business; Computer science	151,568	1.73
Computer science; Materials science	149,538	1.71
Engineering; Environmental science	101,786	1.16

In order to get an impression of the validity of the (automatic) subject reclassifications from *MAG* to *OpenAlex*, we looked at the extreme cases concerning the proliferation of top-level FoSs as given in Table 7. We took random samples of those documents in the common dataset that have a FoS Art resp. Engineering in *MAG*. Additional restrictions were the publication year 2020, the document type *Journal* and the availability of a DOI. Finally, we only chose documents that received a unique top-level FoS in *OpenAlex*. We retrieved the documents and tried to assess the suitability of their FoS assignments in both databases, denoted by a numeric score of 1 (correct), -1 (not correct), and 0 (possibly plausible). Tables 12 and 13 show the respective results, including content information that lead to our assessment and a summed up suitability score for each database. Of course, the applied method can only produce preliminary results with a low accuracy but they could give a hint for further investigations.

Table 12. Assessment of the suitability of unique top-level FoS assignments for a random sample from the 3,717,975 documents assigned to Art in *MAG*

FoS in *OpenAlex*	**DOI of sample document**	**Content information**	***MAG* FoS suitability score**	***OpenAlex* FoS suitability score**
Art	*10.25162/afmw-2020-0005*	Musicology	1	1
Biology	*10.25223/brad.n38.2020.a10*	Botany	-1	1
Business	*10.47287/cen-09844-newscripts*	Chemical & Engineering News: Consumer Products; 2020 holiday gift guide	-1	1
Chemistry	*10.4312/keria.22.1.143-202*	Latin poetry	1	-1
Computer science	*10.1016/s1634-7358(20)44295-0*	Endocrinology (Medicine)	-1	-1
Economics	*10.33008/ijcmr.202019*	Creative performative installation to a film	1	-1
Engineering	*10.1061/(asce)gm.1943-5622.0001816*	Editorial in *Journal of Geomechanics*	-1	1
Environmental science	*10.1002/awwa.1472*	Obituary in *Journal American Water Works Association*	0	1
Geography	*10.4000/geomorphologie.14981*	Archaeology in Special issue on "geomorphologie et environnements karstiques"	0	1
Geology	*10.1180/mgm.2020.16*	Obituary in *Mineralogical Magazine*	1	1
History	*10.18223/hiscult.v9i1.3154*	History and Culture	0	1
Materials science	*10.33383/2020-006*	"Recommendations for restoration of historical transparent coatings in Pushkin Museum"	-1	1
Mathematics	*10.51481/amc.v56i1.823*	Obituary in *Acta Medica Costarricense* (Biomedicine)	1	-1
Medicine	*10.1136/bmj.m3665*	Obituary in *BMJ*	1	1

FoS in *OpenAlex*	DOI of sample document	Content information	*MAG* FoS suitability score	*OpenAlex* FoS suitability score
Philosophy	*10.13125/medea-4529*	Italian Poetry (Cultural Studies)	1	1
Physics	*10.1038/d41586-020-02136-4*	Photo of the sun (News in *Nature*)	-1	1
Political science	*10.1080/10402659.2020.1823575*	*Peace review: A journal of social justice*	1	1
Psychology	*10.1037/amp0000602*	Obituary of a Psychologist	1	1
Sociology	*10.35293/srsa.v37i1.229*	Book review on recent history of South Africa with socio-economic dimension	-1	0
Sum of suitability scores			2	10

Table 13. Assessment of the suitability of unique top-level FoS assignments for a random sample from the 8,472,749 documents assigned to Engineering in *MAG*

FoS in *OpenAlex*	DOI of sample document	Content information	*MAG* FoS suitability score	*OpenAlex* FoS suitability score
Art	*10.22452/sare.vol57no2.10*	Poetry and Fiction	-1	1
Biology	*10.1182/blood.2020008691*	Cells in blood	-1	1
Business	*10.4028/www.scientific.net/amm.896.371*	Tax rules for buildings in *Applied Mechanics and Materials*	1	1
Chemistry	*10.1021/cen-09813-scicon10*	*C&EN* (section Materials) on engineered micro- and nanostructures that mimic spiders	1	1
Computer science	*10.17577/ijertv8is120305*	Comparative study of building rating systems	1	-1
Engineering	*10.1007/s35658-020-0295-y*	Automatized shuttle buses	1	1
Environmental science	*10.3130/aije.85.19*	Study on hygrothermal behavior of wall assembly ... impact of rain penetration and water absorption	1	1
Geography	*10.5632/jila.83.539*	Development of 'businesses-utilizing urban parks'	-1	1
History	*10.1146/annurev-anchem-091119-120456*	Evolution of analytical sciences in the United States: A historical account	-1	1
Materials science	*10.1063/1.5145201*	Development of microLED in *Appl.phys.lett.*	1	1
Mathematics	*10.1007/s11071-020-05950-7*	Obituary in *Nonlinear dynamics*	1	1
Medicine	*10.1055/a-1309-0141*	Dental technology	1	1
Philosophy	*10.1007/s40544-020-0360-9*	Obituary in *Mechanical engineering*	1	-1
Physics	*10.1007/s35658-019-0154-x*	Risk assessment with respect to thermal propagation	1	1
Political science	*10.1038/s41587-020-0702-1*	Correction in *Nature biotechnology*	1	-1
Psychology	*10.12775/jehs.2020.10.05.035*	Future bachelors of motor transport	1	-1
Sociology	*10.4079/gbl.v20.1*	Article published by *Global business languages*	-1	-1
Sum of suitability scores			7	7

Maybe, obituaries and editorials in the sample in Table 12 tend to be classified as Art in *MAG*, but in *OpenAlex*, the scientific subject seems to be more important. Telling by the suitability score, for this sample, the *OpenAlex* subject classifications are definitely more suited. In case of Engineering (see Table 13), there are far more cases which could plausibly be assigned to two or more FoSs, which is also expressed by the same and relatively high suitability scores for both databases.

Level-1 Fields of Study

The FoSs of level-1 are even more interesting for bibliometric evaluations and comparisons because their number is similar to the number of the journal based subject categories in *WoS* and *Scopus* thereby enabling a similar granularity for field normalizations. For our dataset, a total of 74,454 types of reclassifications occur, i.e., on average about 255 for each of the 292 FoSs in *MAG*. Because of their number, it is no longer feasible to present the reclassifications in tabular form. In Figure 4, the most frequent reclassifications of level-1 FoSs from *MAG* to *OpenAlex*, are shown. On this level, a fair amount of symmetry can be detected. For example, in case of Internal medicine, the biggest transfers to other FoSs seem to occur in both directions. This impression remains even if going to a much smaller threshold value as, e.g., 50,000 reclassifications.

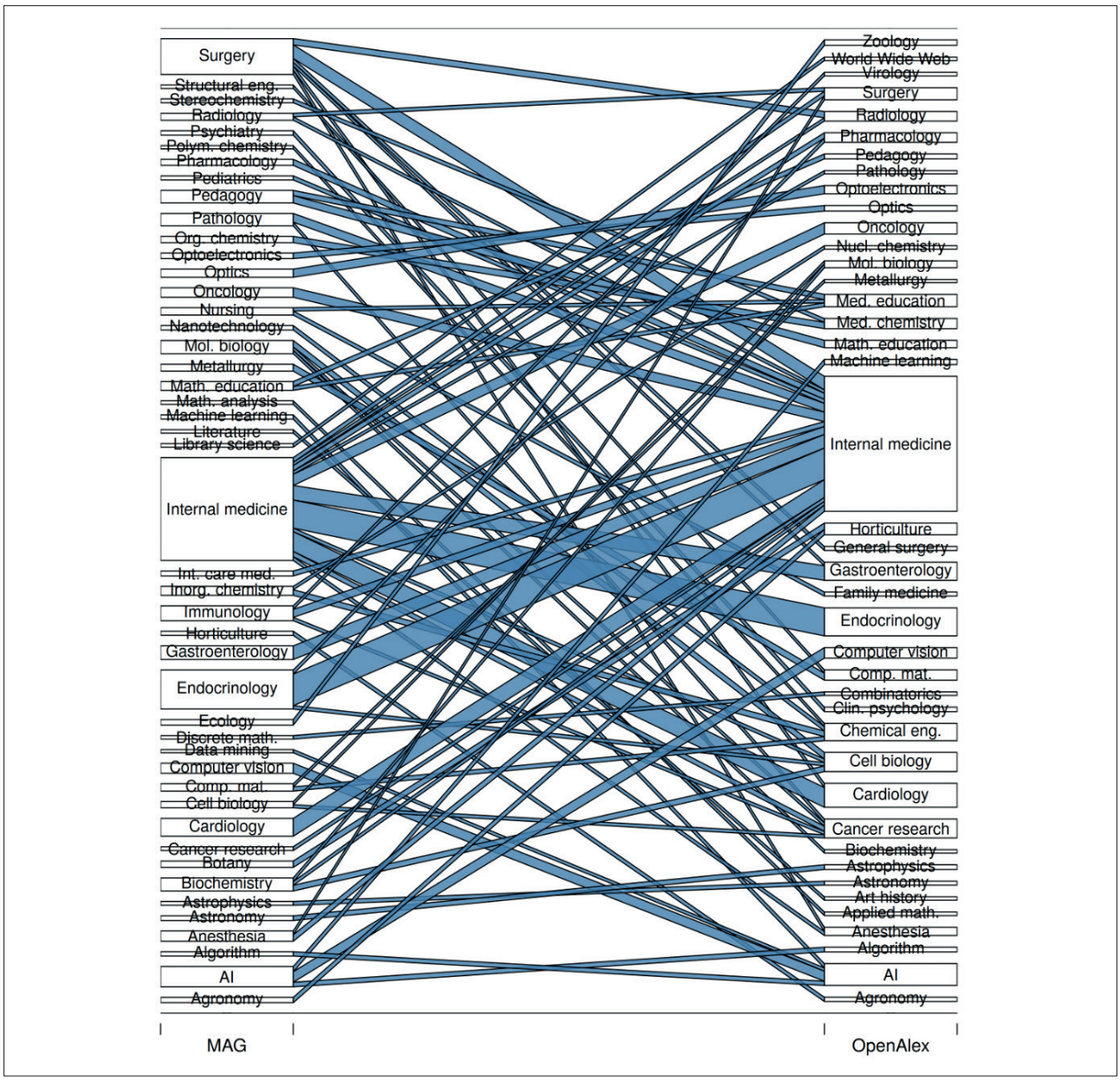


Figure 4. Alluvial diagram for the level-1 FoS reclassifications from *MAG* to *OpenAlex*, showing only reclassifications that occur at least 250,000 times. Some of the original FoS names were rather long and thus abbreviated in this plot. The abbreviated FoS names with their original names can be found in Table A1 in the Appendix.

As mentioned above, eight of the 292 level-1 FoSs present in *MAG* are missing in *OpenAlex*. Their numbers of documents in *MAG* are given in Table 14. Especially, the removal of the two most populated FoSs Analytical chemistry and Pattern recognition with nearly 2 resp. 1.5 million papers in *MAG* is surprising.

Figure 5 displays the numbers of their level-1-FoS assignments in *OpenAlex* occurring at least 30,000 times. Therefore, Ceramic materials is not shown. A first look reveals mostly reclassifications to closely related FoSs.

Table 14. Number of documents in MAG of the eight in *OpenAlex* missing level-1 FoSs

Missing level-1 FoS in *OpenAlex*	#Documents in *MAG*
Algebra	398,751
Analytical chemistry	1,989,709
Calculus	205,406
Ceramic materials	725
Control theory	967,322
Hydrology	524,127
Pattern recognition	1,417,305
Topology	277,926

4. Discussion and conclusions

OpenAlex has transferred practically all works from *MAG* preserving their bibliographic data publication year, volume, first and last page, DOI, and the number of references that are important ingredients for citation analysis.

More than 90% of the *MAG* documents have equivalent document types in *OpenAlex*. Of the remaining ones, especially reclassifications to the *OpenAlex* document types *journal-article* and *book-chapter* seem to be correct and amount to

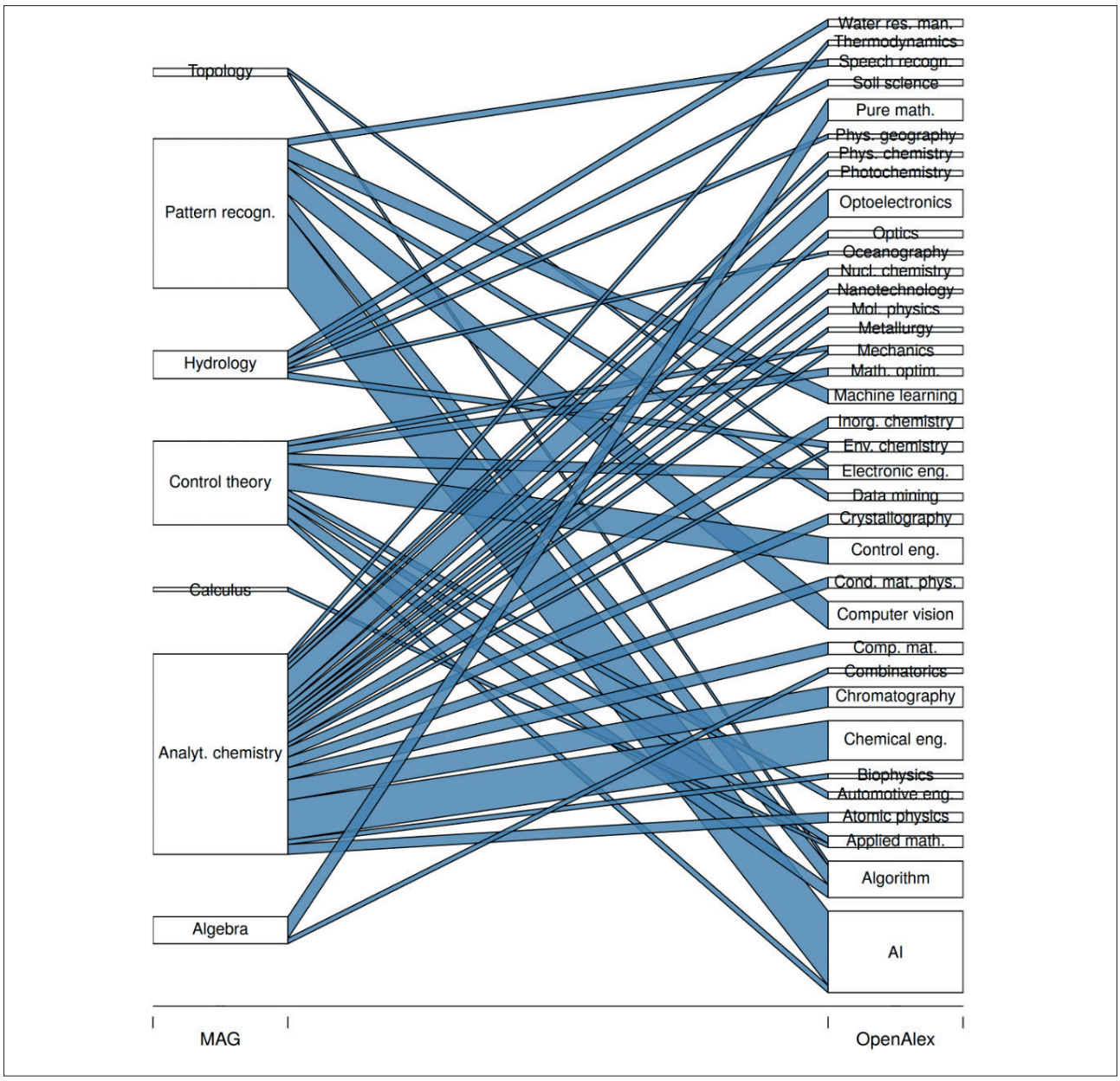


Figure 5. Alluvial diagram for the reclassifications of the eight in *OpenAlex* missing level-1 FoSs from MAG to *OpenAlex*, showing only reclassifications that occur at least 30,000 times. Some of the original FoS names were rather long and thus abbreviated in this plot. The abbreviated FoS names with their original names can be found in Table A1 in the Appendix.

more than 7%, so that the document type specifications have improved significantly from *MAG* to *OpenAlex*. So far, *OpenAlex* seems to be more suited for bibliometric analyses than *MAG*.

As last item of bibliometric relevant metadata, we looked at the paper-based subject classification to FoSs in *MAG* and in *OpenAlex*. We found significantly more documents with a FoS assignment in *OpenAlex* than in *MAG*. On the top level and on level 1, the FoS structure is identical resp. nearly identical, but on the deeper levels the number of available FoSs is drastically reduced to about 10%. A striking feature on the top level is the proliferation and abundant reclassification of the 19 FoSs –very unevenly distributed among them.

This is also the case for a random sample used to assess the suitability of FoS assignments in both databases. On level 1, the reclassifications of FoSs seem to be much more symmetric and the missing eight FoSs to be distributed to closely related ones so that the net effect on the conclusions drawn from previous bibliometric analyses using level-1 FoSs, as by **Scheidsteger** *et al.* (2018), might be small. But that still needs to be investigated as

> Eight level-1 fields of study (Algebra, Analytical chemistry, Calculus, Ceramic materials, Control theory, Hydrology, Pattern recognition, Topology) out of 292 present in *Microsoft Academic Graph* are missing in *OpenAlex*

well as the consequences of *OurResearch* switching to model V2, a different subject classification algorithm, during the year 2022, that promised to bring a substantial improvement (**Priem**; **Piwowar**, 2022).

Overall, *OpenAlex* seems to be at least as suited for bibliometric analyses as *MAG* for publication years before 2021. However, this first impression needs to be checked by further detailed studies.

> In *OpenAlex*, the total number of papers with any field of study is significantly increased and the number of assignments per paper to a top-level field of study is also drastically increased

5. References

Bojanowski, Michał; **Edwards, Robin** (2016). *Alluvial: R package for creating alluvial diagrams*. R package version: 0.1-2. *https://github.com/mbojan/alluvial*

Crossref (2021). *Content type markup guide*. *https://www.crossref.org/documentation/content-registration/content-type-markup-guide*

DOI.org (2019). *DOI System and the ISBN System*. *https://www.doi.org/factsheets/ISBN-A.html*

Harzing, Anne-Wil; **Alakangas, Satu** (2017). "Microsoft Academic: is the phoenix getting wings?". *Scientometrics*, v. 110, n. 1, pp. 371-383. *https://doi.org/10.1007/s11192-016-2185-x*

Kramer, Bianca (2022). *COKI Open metadata report* (Update March 25, 2022). *https://github.com/Curtin-Open-Knowledge-Initiative/open-metadata-report*

Microsoft Blog (2021). *Microsoft Academic*. *https://www.microsoft.com/en-us/research/project/academic*

OpenAlex (2021). *Migration guide*. *https://docs.openalex.org/download-snapshot/mag-format/mag-migration-guide*

OurResearch (2021). *We're building a replacement for Microsoft Academic Graph*. *https://blog.ourresearch.org/were-building-a-replacement-for-microsoft-academic-graph*

Priem, Jason; **Piwowar, Heather** (2022). *OpenAlex-concept-tagging*. *https://github.com/ourresearch/openalex-concept-tagging*

Priem, Jason; **Piwowar, Heather**; **Orr, Richard** (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". In: 26th *International conference on science, technology and innovation indicators (STI 2022)*, Granada, Spain. *https://doi.org/10.5281/zenodo.6936227*

R Core Team (2020). *R: A language and environment for statistical computing*. *https://www.R-project.org*

Scheidsteger, Thomas; **Haunschild, Robin** (2022). "Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020". In: *26th International conference on science, technology and innovation indicators (STI 2022)*, Granada, Spain. *https://doi.org/10.5281/zenodo.6975102*

Scheidsteger, Thomas; **Haunschild, Robin**; **Hug, Sven E.**; **Bornmann, Lutz** (2018). "The concordance of field-normalized scores based on Web of Science and Microsoft Academic data: A case study in computer sciences". In: *23rd International conference on science, technology and innovation indicators (STI 2018)*, Leiden, The Netherlands. *https://hdl.handle.net/1887/65358*

Sinha, Arnab; **Shen, Zhihong**; **Song, Yang**; **Ma, Hao**; **Eide, Darrin**; **Hsu, Bo-June-Paul**; **Wang, Kuansan** (2015). "An overview of Microsoft Academic Service (MAS) and applications". In: *24th International conference on World Wide Web (WWW'15 Companion)*, Florence, Italy. *https://doi.org/10.1145/2740908.2742839*

Visser, Martijn; **Van-Eck, Nees-Jan**; **Waltman, Ludo** (2021). "Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic". *Quantitative science studies*, v. 2, n. 1, pp. 20-41. *https://www.doi.org/10.1162/qss_a_00112*

Wickham, Hadley (2016). *ggplot2: Elegant graphics for data analysis*: New York: Springer-Verlag. ISBN: 978 3 319 24277 4

6. Appendix

Table A1. List of abbreviations of level-1-FoS names used in the alluvial plots Figure 4 and Figure 5

Abbreviated FoS name	Original FoS name
AI	Artificial Intelligence
Analyt. chemistry	Analytical chemistry
Applied math.	Applied mathematics
Automotive eng.	Automotive engineering
Chemical eng.	Chemical engineering
Clin. psychology	Clinical psychology
Comp. mat.	Composite material
Cond. mat. phys.	Condensed matter physics
Control eng.	Control engineering
Discrete math.	Discrete mathematics
Electronic eng.	Electronic engineering
Env. chemistry	Environmental chemistry
Inorg. chemistry	Inorganic chemistry
Int. care med.	Intensive care medicine
Math. analysis	Mathematical analysis
Math. education	Mathematical education
Math. optim.	Mathematical optimization
Med. chemistry	Medicinal chemistry
Med. education	Medical education
Mol. biology	Molecular biology
Mol. physics	Molecular physics
Nucl. chemistry	Nuclear chemistry
Org. chemistry	Organic chemistry
Pattern recogn.	Pattern recognition
Phys. chemistry	Physical chemistry
Phys. geography	Physical geography
Polym. chemistry	Polymer chemistry
Pure math.	Pure mathematics
Speech recogn.	Speech recognition
Structural eng.	Structural engineering
Water res. man.	Water resource management