Tracking content updates in Scopus (2011-2018): a quantitative analysis of journals per subject category and subject categories per journal

Frédérique Bordignon

Tracking content updates in Scopus (2011-2018): a quantitative analysis of journals per subject category and subject categories per journal

Frederique Bordignon[1]

[1] *frederique.bordignon@enpc.fr*
Ecole des Ponts ParisTech, Direction de la Documentation, Champs-sur-Marne, France

Abstract
The aim of this study is to track Scopus content updates since 2011 and more particularly the distribution of journals into subject areas. An unprecedented corpus of data related to sources indexed in Scopus has been created and analyzed. Data shows important fluctuations regarding the number of journals per category and the number of categories assigned to journals. Those variations are very irregular, invisible to the average user and unpredictable over time. They question the reliability of studies based on Scopus data if no precautions are taken. The suggestion is made that category changes should not systematically be applied to all previously indexed publications of a journal, but only to those that will be indexed in Scopus after the new assignment is made.

Introduction

As far as scholarly literature is concerned, two levels of aggregation can be used to delimit scientific areas: the article level and the journal level. Both journals and articles can be classified into fixed sets of subject areas but the delineation of journals at the disciplinary level plays a major role in scientometrics, mainly for analyses based upon the extraction of scientific outputs from databases. That is why those information and reference sources need to be organized through an appropriate and consistent classification scheme (Gómez-Núñez, Batagelj, Vargas-Quesada, Moya-Anegón, & Chinchilla-Rodríguez, 2014). It serves as the basis of profiling authors, research groups, institutions or countries and helps in the making of comparisons and rankings. It is also useful in the calculation of standards for relative citation indicators. And beyond those research evaluation perspectives, journal classifications can also be used in describing the structure of scholarly publication and designing maps of science.

The two following tenets were formulated a long time ago: comparisons should be made in terms of "like with like" (Martin & Irvine, 1983) and over time in terms of fixed journal sets (Narin, 1976). But bibliographic databases do not take these principles into account and the data that is made available is used unsuspectingly by analysts in organizations.

The aim of this study is to track Scopus coverage updates since 2011 and more particularly the distribution of journals into subject areas. Scopus classification of documents is based upon the All Science Journal Classification (ASJC) whose structure does not evolve over time whereas the content of the different categories fluctuates substantially.

Background

Classification for evaluation purposes

(Archambault et al., 2011) stated that no international standard classification scheme exists that supports bibliometric research, and no single classification scheme has been widely adopted by the bibliometric community. Even research funders don't have a standardized classification

system to assess the impact of the funds distributed across different scientific fields (Katz & Hicks, 1995).
Among many initiatives to elaborate efficient classification systems for evaluation purposes, we can mention the following:

- the Steunpunt Onderwijs & Onderzoek Indicatoren (SOOI) implemented for the evaluation unit in Leuven (Glänzel & Schubert, 2003),
- the CHI Research classification (from Computer Horizons Inc) designed for the US National Science Foundation (NSF) (Carpenter & Narin, 1973) and also used by the Canadian Observatoire des sciences et des technologies (OST)
- the Australian Research Council Evaluation of Research Excellence (ERA) classification, abandoned since 2012.

However, it seems that the two most commonly used systems are those on which the Web of Science and Scopus databases are built.

The number and more particularly the diversity of classification schemes complicate comparative analyses (Gómez, Bordons, Fernández, & Méndez, 1996) because of the dual problem of matching categories and delineating journals comparable sets.

Mono vs multi-disciplinary classification systems

Some systems provide a way to classify publications with a great level of details in a restricted research area: for instance the widely used JEL (Journal of Economic Literature) classification system in economics, the Chemical abstracts service in chemistry or the MeSH (Medical Subject Headings) hierarchical system in medicine.
On the opposite, others are appended to multidisciplinary databases that index articles from journals and offer the possibility to retrieve them according to the field(s) the journals are assigned to. Journal level classification systems are of course very convenient but they are known as well to be sometimes too fuzzy, at least not so accurate as article level classifications. Indeed it is well known that most journals contain articles dealing with a relatively broad range of themes, in spite of their "main subject". Thus, a subject delimitation based on journal classification will probably contains some articles weakly related with the target subject, while some pertinent articles will be missing (Bensman, 2001; Gómez et al., 1996). And (Pudovkin & Garfield, 2002) said about the Web of Science classification system that "journals are assigned to categories by subjective, heuristic methods. In many fields these categories are sufficient but in many areas of research these ‘classifications’ are crude and do not permit the user to quickly learn which journals are most closely related."

Multiaffectation classification systems

Another limit of journal level classifications is due to the fact that many journals are assigned to multiple categories to better represent the scientific themes their articles deal with. A mutually exclusive classification is of course more convenient, in particular because it prevents a journal to being counted more than once. What is more, those classifications are generally not well documented (Archambault et al., 2011), and therefore there is no indication about why one or more categories were chosen for a particular journal. (Wang & Waltman, 2016) found that a significant share of the journals in both databases, but especially in Scopus, seem to have assignments to too many categories and then suggested to adopt a stricter policy supported by the use of citation analysis when assigning journals to categories.

Multiaffectation is supposed to reflect interdisciplinarity but in the end, multidisciplinary journals (eg: *Nature*, *PNAS*, and *Science*) are the most poorly managed and that is what leads (Wang & Waltman, 2016) to reconsider journal classification systems at a more fundamental level and warn an increasing share of publications cannot be properly classified at the journal level because of the increasing popularity of large multidisciplinary journals (eg: *PLoS ONE*). Scopus allows the assignment of a journal to several ASJC categories.

Stability over time

Exploring the limits of existing classification schemes and trying to improve them has given rise to many studies in the field of bibliometrics and scientometrics. But among them, the problem of the changes over time has more rarely been tackled and assessed. At least, the problem of fitting new journals into existing schemes has been dealt with: (Leydesdorff, 2002) with the idea of comparing structural changes in a database with reorganizations of relations among previously included journals concludes that "if one does not systematically account for redelineation in the groupings over time but uses "fixed journal sets" instead, one risks making a prediction of performance with reference to an outdated unit.". Despite this conclusion, Scopus (more particularly the possibility to request for Scopus data according to preset corpus of journals in different subject areas) is still the easiest way to retrieve data to produce reports in many organizations.

The question is whether these analyses are reliable when they cover different periods of times and subject areas if the different sets of journals are not stable and if the changes are not clearly reported. Indeed, queries in Scopus do not take into account any journal assignation update according to the publication year the query is based upon.

In this study, we investigate 2 kinds of potential changes in Scopus: (1) number of journals per categories (2) number of categories per journal, both impacting the delineation of categories and therefore the data retrieved from Scopus.

Methods

The All Science Journal Classification scheme

Scopus journal classification system is called the All Science Journal Classification (ASJC). There seems to be no official description about the way it is constructed. It has always been freely available online either from a dedicated page on the Elsevier website and an Excel file available for download, or from the former JournalMetrics website. It can also been downloaded from Scopus database (*Browse sources* page).

It is commonly described as consisting of two levels, but there is actually a third level above all, differently called Top-Levels, Supergroups or Subject areas (depending on time periods and downloadable files). This uppest level is not used at all in Scopus but can be used to filter out journals in the Excel file. The lowest level has 307 subfields, and the intermediate level includes 26 fields called *Subject areas* in Scopus. There is another field and another subfield for the *Multidisciplinary* category. All the subfields are assigned a 4-digit code.

We do not know how the assignment of journals to fields and subfields is decided but we can infer it is done by the Scopus Content Selection and Advisory Board, "an international group

of scientists, researchers and librarians who represent the major scientific disciplines" (Elsevier website, 2019[1]).

Table 1. ASJC journal classification system

Supergroups	*Fields*	*No. of Subfields*
-	Multidisciplinary	1
Health Sciences	Medicine	48
Health Sciences	Nursing	23
Health Sciences	Veterinary	4
Health Sciences	Dentistry	6
Health Sciences	Health Professions	16
Life Sciences	Agricultural and Biological Sciences	11
Life Sciences	Biochemistry, Genetics and Molecular Biology	15
Life Sciences	Immunology and Microbiology	6
Life Sciences	Neuroscience	9
Life Sciences	Pharmacology, Toxicology and Pharmaceutics	5
Physical Sciences	Chemical Engineering	8
Physical Sciences	Chemistry	7
Physical Sciences	Computer Science	12
Physical Sciences	Earth and Planetary Sciences	13
Physical Sciences	Energy	5
Physical Sciences	Engineering	16
Physical Sciences	Environmental Science	12
Physical Sciences	Materials Science	8
Physical Sciences	Mathematics	14
Physical Sciences	Physics and Astronomy	10
Social Sciences	Arts and Humanities	13
Social Sciences	Business, Management and Accounting	10
Social Sciences	Decision Sciences	4
Social Sciences	Economics, Econometrics and Finance	3
Social Sciences	Psychology	7
Social Sciences	Social Sciences	22
4 supergroups	**26 fields**	**307 subfields**

The *Multidisciplinary* field and its unique subfield is dedicated to journals with a very broad multidisciplinary scope like *Nature*, *Science* or *Scientific reports.*

In all fields, there are 2 quite similar subfields:

- one whose label starts with the "*General*" mention and code ends with 00,
- the other whose label ends with the "*(miscellaneous)*" mention and code ends with 01.

It is impossible to say what led the Scopus experts panel to choose between the *General* subfield or the *Miscellaneous* corresponding subfield. There are many examples of journals assigned to both (for example, *Biology Letters*, assigned to the "*General Agricultural and Biological Sciences*" subfield and the "*Agricultural and Biological Sciences (miscellaneous)*" subfield.

Like other classification schemes, the ASJC has been criticized, most frequently because of confusing subfield labels (*Linguistics & Language* and *Language & Linguistics*, (Wang & Waltman, 2016)) or strong imbalanced distribution of journals and therefore documents among

[1] https://www.elsevier.com/solutions/scopus/how-scopus-works/content#content-policy-and-selection, available online on February 2019

the fields (Jacsó, 2013). There have been attempts to improve it (Gómez-Núñez et al., 2014; Jacsó, 2013) but still few considerations about the impact of coverage changes over time.

The nomenclature structure itself is stable since 2011, no new field or subfield has been created over the period we are interested in. Codes remained the same and names of fields and subfields as well, excepted for the *General* field which changed name in 2016 and was renamed *Multidisciplinary* and all the subfields ending with the "*(all)*" mention (eg: *Agricultural and Biological Sciences (all))* that changed name and have been started with "*General*" since 2017 (eg: *General Agricultural and Biological Sciences*).

Data

We retrieved all the title list files Elsevier has released twice a year since 2011 to investigate what content is included in Scopus. Journals, trade journals, conferences and book series are listed but we only focus on journals in this study. Most of those files are still available online thanks to the Wayback machine website. They are the best way to retrieve the metadata needed to our study. We only kept one file a year (the one published at the end of each year) and compiled the 8 files into a single dataset. The aggregated data (Bordignon, 2019) used for this study is available for reuse and further investigation (SNIP values and Open Access status are included in the dataset even if they are not analyzed in our study).

Results

Inclusion and withdrawal of journals at the category level

Table 2. Number of journals included in Scopus and annual growth

	2011	2012	2013	2014	2015	2016	2017	2018
No. of journals	28 335	29 561	31 154	32 332	33 058	33 810	34 772	36 189
Annual growth	-	+4,3%	+5,4%	+3,8%	+2,2%	+2,3%	+2,8%	+4,1%

As far as Scopus content is concerned, the most significant change since 2011 is the increasing number of journals indexed in the database (+28% between 2011 and 2018, with most important increases in 2012 (+4,3%) and 2013 (+5,4%)). Very few journals are merely dropped (min=27;max=318). And even inactive journals are sometimes added to the index (inactive either because they changed name, merged with another journal, splitted or simply ceased to publish anything).

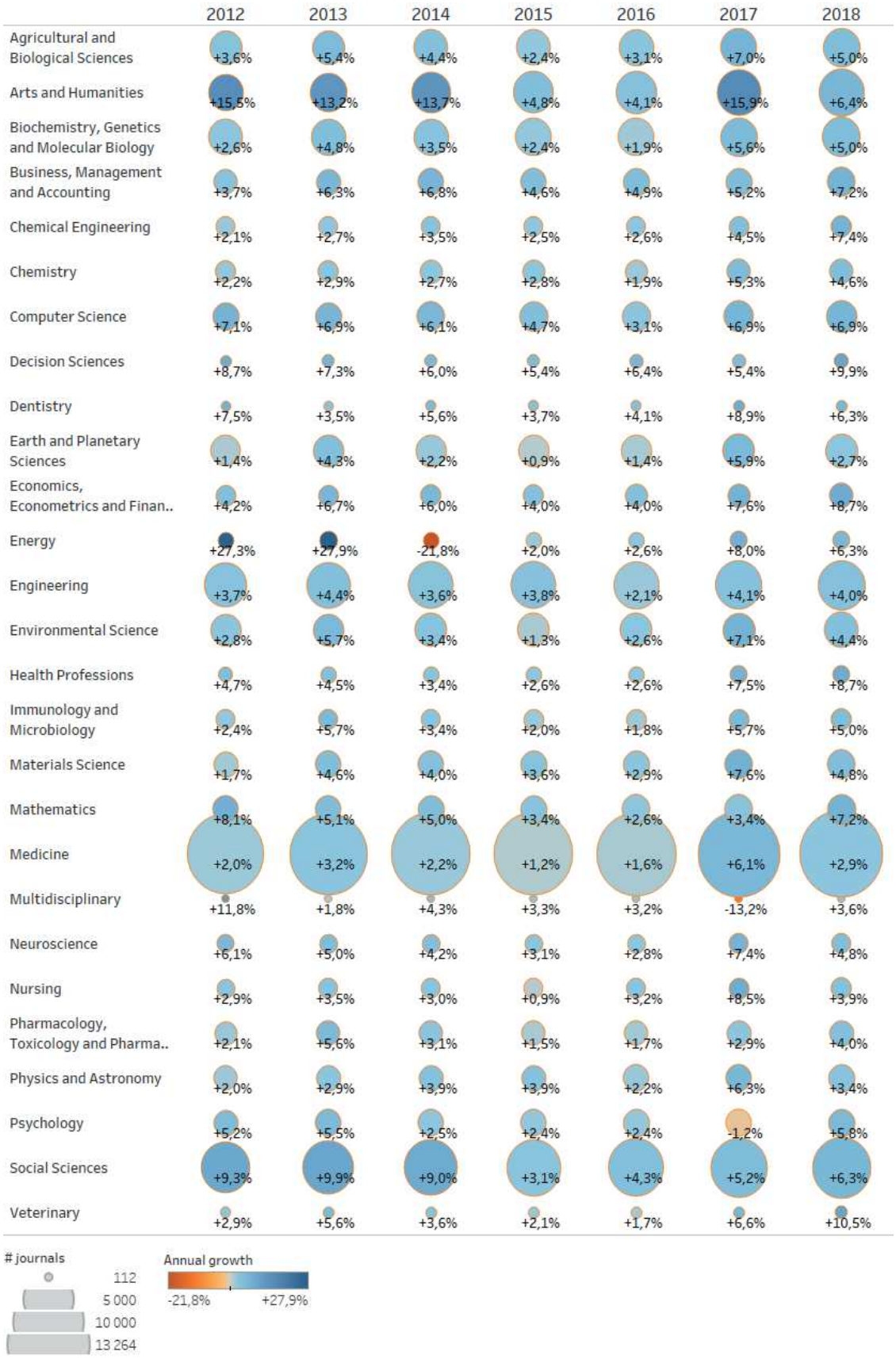

Figure 1. Number of journals per field and annual growth

The evolution of the number of journals is contrasted from one field to another. Here are the highlights Figure 1 reveals:

- in 2012, 2013 and 2014, among the largest fields (2000+ journals), the fields *Arts & humanities* and *Social sciences* had the highest increase (from 9% to 15.5% annual growth). Another large inclusion of sources also occurred in 2017 in *Arts & humanities* (+15.9%, ie: 506 journals added)
- in general, there has been no significant increase in any field in 2015 and 2016, with a maximum of +5.4% in 2015 and a maximum of +6.4% in 2016 (both concerning the *Decision sciences* field).
- in 2017, 500+ journals were added to the *Arts & humanities* field (+15,9%)
- few fields are undergoing decreases:
    - *Psychology* in 2017, -1.2% but this only represents 15 journals, 73 were finally added the following year
    - *Multidisciplinary* in 2017 also with -13.2%, but this amounts to only 17 dropped journals
    - the *Energy* field must be considered as a particular case: indeed it recorded both the highest increase and the highest decrease over the entire period, with the addition of 126 journals in 2013 and the same amount of sources withdrawn the year after. Out of the 126 journals added in 2013, 91 were removed from Scopus in 2014 (all belonging to the *General Energy* subfield).

These updates are unpredictable and have inevitably an impact on comparative studies that are conducted on those fields at different periods of time.

Apart from these fluctuations (mainly additions) in the number of journals per field, it is important to know whether these additions are newly included journals or whether they "come from" other fields, in other words whether journals would change field/subfield, be assigned to more fields/subfields or withdrawn from any field/subfield.

Table 3. Annual percentage of journals whose assignment to fields and subfields has been updated

2012		2013		2014		2015		2016		2017		2018	
subfield	field	subfield	field	subfield	field	subfield	field	subfield	field	subfield	field	subfield	field
0,1%	0,1%	0,2%	0,2%	1,0%	1,0%	0,5%	0,5%	0,3%	0,3%	**10,6%**	**6,9%**	1,5%	1,0%

Table 3 clearly shows that field or subfield shifts are very unusual. This means that a journal is only very rarely reassigned to more or less fields/subfields. The changes mentioned above are therefore only due to additions or withdrawals of journals from the Scopus index. However, a significant change can be observed between 2016 and 2017; Table 4 focuses on this time period and shows that, apart from the *Multidisciplinary* field (and only 33 sources involved), the field for which most shifts are detected is *Psychology* with a significant share (21.6%) being reassigned to more or less subfields.

Table 4. Number and percentage of journals per field with subfield shifts between 2016 and 2017

Fields	No. of journals with subfield shifts	% of all journals in the field
Multidisciplinary	33	25,60%
Psychology	274	21,6%
Immunology and Microbiology	132	18,0%
Biochemistry, Genetics and Molecular Biology	439	17,7%
Health Professions	85	17,7%
Pharmacology, Toxicology and Pharmaceutics	172	16,3%
Neuroscience	99	15,9%
Veterinary	33	13,60%
Medicine	1630	13,4%
Nursing	88	12,5%
Environmental Science	239	12,1%
Social Sciences	652	11,2%
Engineering	376	9,5%
Computer Science	136	8,7%
Earth and Planetary Sciences	149	8,1%
Agricultural and Biological Sciences	186	8,0%
Arts and Humanities	201	6,3%

The Scopus search interface does not allow to query or filter on subfields. But skilled analysts who use the source list file can do this sorting after having exported the bibliographic data; their comparative analyses are likely to be biased because of too important changes between 2016 and 2017.
In addition, some world university rankings by subject are based on citation indicators collected across several subfields. The results of these rankings are necessarily skewed by these significant updates in Scopus.

Number of categories per journal

Our 2018 data shows that the maximum number of fields assigned to a journal is 9 (*The Bulletin of mathematical biophysics*) whereas the highest number of subfields is 13, assigned to *Journal of Geophysical Research*. This example reveals several technical problems actually: first of all, this journal is organized in 7 disciplinary sections (eg: *JGR: Atmospheres*, *JGR: Biogeosciences* etc.). Those sections are not integrated into the Elsevier title list, this might be the reason why so many subfields are assigned to this source. On the other hand, when querying Scopus sources index about *Journal of Geophysical Research,* there are 2 answers: one for the *stem* journal without any mention of sections, the other for a single specific section (*Solid Earth*). And finally, when searching for any documents with *Journal of Geophysical Research* as source title, relevant results indicate the complete correct titles of all the discipline sections. Even if further investigation is needed to measure the extent of the issue, it seems that the 3 sources of information about Scopus content are not consistent.

As far as our dataset is concerned, unsurprisingly, it shows an increasing average number of fields (+1,49% since 2011) and subfields (+3,11% since 2011) assigned to journals, which seems to attest the increasing interdisciplinarity of science (Morillo, Bordons, & Gómez, 2003). Our calculation of the average number of fields assigned to journals is consistent with (Wang & Waltman, 2016) results (2.1 in Scopus) as shown in Table 5.

But consistently with what we stated earlier, this increase is not due to updates at the journal level but almost exclusively due to the addition of journals to the database index. Since 2013, those newly included sources have always been assigned to more fields and subfields on average than those previously indexed.

Table 5. Average number of fields and subfields per journal for added or previously included ones

		2011	2012	2013	2014	2015	2016	2017	2018
Avg number of fields	Added	1,48	1,51	1,56	1,57	1,61	1,66	1,67	1,66
	Previously included	1,53	1,53	1,53	1,53	1,53	1,54	1,54	1,54
Avg number of subfields	Added	1,91	2	2,02	2,08	2,26	2,37	2,45	2,38
	Previously included	1,99	1,98	1,98	1,99	2	2,01	2,01	2,03

This is a global tendency and further studies will be able to reveal differences among fields. It should also been worth checking whether newly created journals (and not newly included ones) are being assigned more fields and subfields as well.
It seems unlikely that interdisciplinarity will only arise on newly added journals. This means that the journals already in the database should be re-examined by the Elsevier experts panel. And of course this reinforces the idea that indexing at the article level better reflects reality.

Discussions and perspectives

Whether analysts work directly in Scopus or use the data Elsevier makes available in downloadable files, they cannot perform the time-consuming analysis work that would assess if updates to Scopus coverage are not too substantial and if comparing reports produced from one year to the next is still possible.

Moreover, the large volume changes we have highlighted in some categories certainly do not reflect the scientific reality of the field but rather Elsevier's objectives to increase its coverage. And yet, the consequences can be significant: for example, on SNIP values due to an unstable scope of journals and therefore a very unstable citations rate, or on international thematic university rankings whose evaluation criteria are based partly on the collection of citations and outputs according to subject areas.

Without giving up the extension of the coverage and the necessary updating of the database, Elsevier should inform the user of Scopus content updates in order to prevent potential impacts on the resulting analyses. This is obviously something very complex to set up in the interface, but one cannot assume that all users regularly consult the title list file. One possibility is to reflect category changes of a journal only on newly added publications (recently published or not) and not on all publications already present in the database. It will mitigate the bias for university rankings or the calculation of indicators.

As for the increase in the average number of fields and subfields per journal, since it is limited to additions, it cannot be said that it can be used to support work on interdisciplinarity.
On this particular point, it should be reiterated that a journal can be added to the list of indexed sources even if it has a long-established history, or even if it is inactive. Therefore, there is a

limit to our analysis since we would have to examine whether the increase in the average number of fields/subfields per journal is true for all journals added to the index or more particularly for those created and included at the same period of time.

Conclusion

We know that classifications at the article level are more relevant, but since bibliographic databases offer the possibility of queries, analyses and data exports based on the classification of journals, it is important to know to what extent this could impact their reliability for bibliometric analyses.

We revealed the existence of very important updates in the Scopus database which can have a significant impact, depending on the scope of the analyses carried out. We also showed that these fluctuations were very irregular, invisible to the average user and unpredictable. That is why we suggested that category changes should not systematically be applied to all previously indexed publications of a journal, but only to those that will be indexed in Scopus after the new assignment is made.

References

Archambault, É., Beauchesne, O. H., Caruso, J., Archambault, É. ;, Beauchesne, O. H. ;, & Archambault, C. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th international conference of the international society for scientometrics and informetrics. E.C.M. Noyons, P. Ngulube, J. Leta (Eds.)* (pp. 66–77). Retrieved December 23, 2018, from www.sciencemetrix.com

Bensman, S. J. (2001). Bradford's Law and Fuzzy Sets: Statistical Implications for Library Analyses. *IFLA Journal*, *27*(4), 238–246.

Bordignon, F. (2019). *Scopus sources title list: aggregated data*. *Mendeley Data*.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, *24*(6), 425–436.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*(2), 223–235.

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369–383.

Jacsó, P. (2013). The need for end-user customization of the journal-sets of the subject categories in the SCImago Journal Ranking database for more appropriate league lists. A case study for the Library & Information Science field. *El Profesional de la Informacion*, *22*(5), 459–473.

Katz, J. S., & Hicks, D. (1995). The Classification of Interdisciplinary Journals: A New Approach (Version 2.0). *Proceeding of The Fifth Biennial Conference of The International Society for Scientometrics and Informatics, Rosary College, River Forest, Il, USA, June 7-10, 1995*. IEEE.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology*, *53*(12), 987–994.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinary in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249.

Narin, F. (1976). *Evaluative Bibliometrics: the use of publication and citation analysis in the evaluation*

of scientific activity. *Computer Horizons, Inc Project No. 704R - Contract report to the National Science Foundation*.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113–1119. John Wiley & Sons, Ltd.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364.