Keyword occurrences and journal specialization

Gabriele Sampagnaro[1]

Abstract

Since the borders of disciplines change over time and vary across communities and geographies, they can be expressed at different levels of granularity, making it challenging to find a broad consensus about the measurement of interdisciplinarity. This study contributes to this debate by proposing a journal specialization index based on the level of repetitiveness of keywords appearing in their articles. Keywords represent one of the most essential items for filtering the vast amount of research available. If chosen correctly, they can help to identify the central concept of the paper and, consequently, to couple it with manuscripts related to the same field or subfield of research. Based on these universally recognized features of article keywords, the study proposes measuring the specialization of a journal by counting the number of times that a keyword is Queryrepeated in a journal on average (Sj). The basic assumption underlying the proposal of a journal specialization index is that the keywords may approximate the article's topic and that the higher the number of papers in a journal based on a topic, the higher the level of specialization of that journal. The proposed specialization metric is not invulnerable to a set of limitations, among which the most relevant seems to be the lack of a standard practice regarding the number and consistency of keywords appearing in each article.

Keywords Journal specialization · Keyword occurrences · Interdisciplinary

Introduction

This paper aims to contribute to the debate on the specialization and compartmentalization of academic journals and, indirectly, on their degree of interdisciplinarity. An interdisciplinary journal is an academic publication that publishes articles and reviews covering themes and approaches from different disciplines (Augsburg, 2016). These journals seek to overcome the boundaries of individual disciplines and promote collaboration and integration of research approaches from different fields. Articles published in interdisciplinary journals often address complex issues that require understanding and solutions that cannot be provided by a single discipline. Interdisciplinary journals focus on promoting research and academic debate that goes beyond the limits of individual disciplines and promotes a

✉ Gabriele Sampagnaro
  gabriele.sampagnaro@uniparthenope.it

1   Department of Business and Quantitative Studies, University of Naples Parthenope, Naples, Italy

more holistic approach to analysing and understanding problems. These journals are often sought out by those looking for a broader perspective on a topic or issue and are used as a source of information for interdisciplinary research and academic projects. In contrast to interdisciplinarity, we have journals with a strong degree of disciplinarity, which we identify here with the concept of specialization. A specialized journal is an academic source whose contributions are strongly concentrated in a field bodies of knowledge or research practice. The specialization level and interdisciplinarity of journals may be considered complementary terms: the higher the interdisciplinarity of a journal is, the lower the level of specialization and vice versa.

This paper intends to contribute to developing a debate about the disciplinarity level of journals by introducing an index of journal specialization (Sj index).

The literature on the degree of disciplinarity of journals is scant in formal terms, but it is extensive in substantive terms. In fact, if we consider specialization as a complementary term to interdisciplinarity, it is straightforward to count the bulk of studies on interdisciplinarity as indirect references to the topic of journal specialization.

From the analysis of the literature on interdisciplinarity, it emerges that while the attention on interdisciplinary research has grown in recent decades (Leydesdorff, 2007; Glänzel & Debackere, 2022), the academic literature on this topic has shown a lack of consensus on the most appropriate measurement approach for interdisciplinary research (Leydesdorff & Rafols, 2011; Wang & Schneider, 2020). Moreover, the categorization of scientific articles and journals by topic is one of the most difficult and essential problems of information science.

Specifically, and in a quantitative context, there are at least two different methods to classify the interdisciplinarity of research: journal-level and paper-level categorization. The debate on which of these two methods of categorization provide the most appropriate solution is still open (Abramo et al., 2018; Milojević, 2020). The most common research categorization approach is the journal-level approach, based on the assumption that journal subject categories (SCs) identify disciplines (Carpenter & Narin, 1973, Zwanenburg et al., 2022).

However, using an a priori discipline structure is not invulnerable to criticism. The unambiguous categorization of journals through subject categories appears to be rare due to the fuzziness of journal sets (Bensman, 2001). Further, Journal-level categorization cannot assign a single or few categories to interdisciplinary journals (such as Nature or Science), ad other concerns about the risk of disambiguation have been raised.[1]

To overcome these limitations, it seems reasonable to introduce a paper-level categorization approach based on one of the most granular items in an article, such as the title words, as suggested in their final remarks by Eck and Waltman (2012) or keywords, as proposed by this study.

This paper tries to produce a measure of specialization (and therefore indirectly of interdisciplinarity) of journals using paper-level categorization that uses research keywords as an element of analysis. Article keywords are generally considered a crucial element in

---

[1] Adopting a single SCs system of classifying disciplines (including the WoS-SC classification) may tend to misclassify journals by lumping unconnected ones into single and erroneous subject categories (Milojević, 2020). The misclassification of papers caused by the journal-level method of discipline classification may be severe since a 50% accuracy rate in assigning papers to specific research has been observed (Shu et al., 2019). Furthermore, erroneously categorizing research through a journal-level approach may generate an inconsistent research evaluation output (Shu et al., 2020).

understanding the contents of an academic paper (Callon et al., 1986; Choi et al., 2011; Hartley & Kostoff, 2003; He, 1999; Whittaker, 1989).

As these keywords summarize and represent the paper's topic, they represent carefully selected terms that are deemed essential and significant. This is because scholars can quickly identify papers by scanning through article keywords, which are known to consist of the most important terms that represent papers. Additionally, research keywords can reveal related terms that may have previously been unknown and expand search queries.

The use of keywords seems to be one of the most effective ways authors and editors enhance the chances of potential readers finding their articles. Since keywords are not limited to predetermined categories, as they can be freely selected by the authors, they permit readers to make certain that the research article is relevant and support editors and research databases in enhancing the relatedness of articles on a specific topic (Hartley & Kostoff, 2003). Furthermore, keywords selected by authors have been shown to be useful for understanding a discipline (Li, 2018; Onyancha, 2018; Xu et al., 2018), even though Tsai et al. (2011) have shown that expert authors tend to select keywords that better represent the content of their article compared to those chosen by novices.

Moreover, it has also been found that some statistical characteristics of keywords (e.g., number of keywords and percentage of new keywords) have significant relations with citation counts (Uddin et al., 2016).

Based on the assumption that keywords are able to identify the content of the article, the approach proposed in this study identifies the frequency of keywords in a journal as a possible indicator of its specialization: the higher (*the lower*) the occurrences of the keyword among articles published by a single journal, the higher its discipline specialization (*its interdisciplinarity*).

By using a dataset consisting of 88,583 articles published in 50 journals, we propose the use of a specialization index, Sj, interpreted as the average frequency at which a keyword appears in the j_{th} journal.

The main peculiarity of this index is its extreme simplicity and replicability due to the availability of freely accessible bibliometric software (i.e., VOSviewer, Eck & Waltman, 2010) that quite easily collects keywords from a dataset of papers.

However, analysing the keyword dynamics requires caution in several aspects. The number of keywords collected in a dataset of articles can be correlated with many factors, such as the number of articles collected, the presence or absence of a publisher's policy that requires the inclusion of a minimum or a maximum number of keywords, the degree of specialization of the article, and the degree of attention given by a researcher to the selection of their keywords. The analysis of the dynamics of keywords still seems to be an unexplored theme, while their use as a possible key factor to screen the specialization of journal research is a novelty.

The essential contribution of this study is to enhance the literature related to publication-level classification systems by introducing a new method of investigation about the specialization of academic journals based on the resource of keywords in a journal that attempts to approximate the level of concentration of articles on a specific topic. By assuming that keywords approximate the article's topic, this study suggests that the higher the occurrences of keywords, the higher the level of specialization of that journal.

The remaining paper is structured as follows: Sect. "Related literature" provides context for the proposed measure by comparing it to other journal disciplinarity/interdisciplinarity measures. Sect. "Methodology" reports the methodology of the specialization index. Sect. "The dataset" describes the way in which the sample of journals was composed. Sect. "Keyword dynamics across journals." describes some unobserved dynamics

of keywords of journals, while Sect. "Results" reports the calculation of the metrics for the identified sample. Finally, Sect. "Conclusions and limitation" concludes and outlines the implications of this study.

Related literature

To our knowledge, this is the first study that attempts to propose a metric of specialization by using author keywords. At the same, time it is not the first study that propose an index of specialization. Boyack and Klavans (2011) propose to measure the journal specialization (indicated with the term "journal specificity") by introducing more than one quantity including a textual coherence indicator that is the most close to the index presented in this manuscript. By leveraging the title and abstract information of papers, the authors employed word probability vector techniques to generate clusters and structural insights. Specifically, the index of journal specificity proposed is based on the Jensen–Shannon divergence since it measures the similarity (and then the divergence) between two probability distributions: the likelihood of occurrence of a word in a document (an article) and the likelihood of occurrence of the same word in a Journal (where the article has been published). This approach enables to capture meaningful data about the scientific domain of academic article and it follows other studies in which quite similar text analysis have been used (Boyack & Klavans, 2010; Braam et al., 1991; Glänzel & Czerwon, 1996; Jarneving, 2001).

By comparing the Boyack and Klavans (2011) metric with the *Sj* index, it becomes apparent that while *Sj* offers insights into the frequency of specific keywords used within a specific journal, the other index determines the textual coherence of individual journals by comparing word probability distributions. In summary, these two metrics differ both in terms of the data used (keywords versus title and abstract) and the methodological approach (probability of occurrence versus actual occurrence).

Other studies discuss, in some case partially, of journal specialization even though without a proposal of metric. The concept of journal specialization has been discussed by Glänzel et al., (1999) who recognizes that the demarcation of subject areas through journal assignment is inherently less precise compared to using subject headings from individual publications. In these circumstances, keywords can be the considered as a flexible means of tracing the dissemination and trajectories of knowledge, as they are highly indicative of the concepts and subjects covered in articles (Xu et al., 2018). By using keywords extracted through an automatic method (Frantzi et al., 2000[2]), rather than author's keywords as reported in our study, Xu et al. (2018) examine the formation of interdisciplinary knowledge through the lens of keyword evolution. By doing so, it is possible to gain insights into the specific developmental characteristics of interdisciplinarity, which provide a potential timeline for the interdisciplinary formation process. Specifically, the formation of an interdisciplinarity approach around a specific topic of research progresses through some significant phases (a latent phase, an embryo phase, and a mature phase) and this evolution may be captured through the evolution of keywords related to a topic in different domains of science. In summary, while keywords may be considered as a valuable method to analyze the concentration of subjects and concepts inside a single journal, if collected in different journals and referred to a specific subject, they

---

[2] Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. International Journal on Digital Libraries, 3(2), 115–130.

can also be considered tools for analysis of the interdisciplinarity attitude of that specific subject and also to predict knowledge evolution (Choi et al., 2011).

Griffiths and Steyvers (2004) introduce a statistical inference algorithm for Latent Dirichlet Allocation (LDA), a generative model that considers documents as a combination of topics. The implementation of this algorithm allows them to explore the topic dynamics and the identification of words' significance in the semantic content of documents. This results is in part related to our study to the extent that it represents a valuable example of how to calculate the frequency of topic in a cluster of documents (including academic journals). In any case, the most distinguished characteristic of the *Sj* Index when compared with other explicit or implicit indicators of journal specialization rely on the circumstances that it is not based on probability calculations neither in a more complex algorithm while it represents a metric expressed by a deterministic process.

Methodology

The content of a paper may be described by many features, such as title, abstract, keywords, and discipline classification coding. Keywords represent a "subject heading" that should help readers understand the central concept of the paper and its fields of concern (Hartley & Kostoff, 2003).

Based on this attitude, the underlying rationale of the recourse to keywords to analyse the topic specialization of a journal is that the more times a keyword is replicated in a journal, the higher the number of papers related to the same or similar concepts.

To formalize this idea, we indicate with K_j (with $K_j > 0$) the total number of unique keywords contained in a set of *m* articles of journal *j*, and with OCC_{ij} (with $OCC_{ij} > 0$ and where $i = 1,..,K_j$) the occurrences of each i_{th} unique keyword appearing in journal *j*. The term "unique" means that, for example, if $K_j = 10$, there are *m* articles of journal j containing 10 keywords, each of which is replicated OCC_{ij} times.

Since OCC_{ij} represents the number of times a unique keyword is found in a single journal *j* and since keywords are items unique to each paper, OCC_{ij} should be positively related to the number of articles focused on a specific field of research. In other words, it is possible to consider OCC_{ij} as the number of papers that use the i_{th} keyword in the j_{th} journal.

Based on this specification, we propose measuring the specialization of the j_{th} journal through the level of density of keywords *Sj*, which can be written in the following form:

$$S_J = \frac{\sum_{i=1}^{K} OCC_{ij}}{K_j} = \frac{OCC_j}{K_J} \tag{1}$$

where OCC_{ij} and K_j have already been defined, while *OCCj* represents, by construction, the total number of keywords (including duplications) selected by all authors of journal *j*.[3]

---

[3] Suppose that a journal has only the following three author keywords: *red*, white, *green*. Then, suppose that the keyword red has an occurrence of 4 (i.e., there are four papers that use this keyword), white has an occurrence of 5, and *green* has an occurrence of 6. In these conditions OCC_j is equal to 15 (4+5+6), and $S_j = 5$ (15/3), meaning that, on average, a keyword is repeated 5 times in the journal. The higher this number, the higher the number of papers in the journal on the same topic. To clarify further the concept of OCC_j and S_j, suppose that the journal has now also a fourth keyword, blue, and that this keyword has an occurrence of 1 (i.e., the keyword blue is present in only one paper). In these new conditions, OCC_j is equal to 16 (4+5+6+1), while S_j is equal to 4 (16/4, where the denominator is the number of unique keywords that passes from 3 to 4).

By expressing the index in this way, *Sj* appears to be immediately interpreted as the number of times that a generic keyword on average appears in the j_{th} journal.

The simplicity of the index and its immediate comprehensibility represent its primary properties.

However, as discussed in the remainder of the paper, the level of *Sj* may be affected by a special link between OCC_j and K_j. Basically, when someone compares the number of unique keywords appearing in journal j (K_j) with the total number (including duplicates) of keywords selected by authors of journal j (OCC_j), it is reasonable to expect that OCC_j increases in *Kj*. This relationship could be explained by probability theories. In particular, assume that selecting a keyword is a mechanism analogous to drawing a card from a deck of *W* cards with replacement, where *W* is the (unknown[4]) size of the vocabulary from which researchers choose their keywords and where any drawing is independent of the others. Under these circumstances, the number of times a researcher selects a keyword (i.e., takes a card from the deck) already chosen by another researcher should theoretically be approximated by the binomial distribution $X \sim B(n, p)$ with parameters $n \in W$ and $p \in [0,1]$, where the probability of success is positively related to the number of trials. In other words, the positive link between the number of keywords in a journal (K_j) and the number of times a keyword may be selected more than one time in that journal (OCC_{ij}) is because with a larger number of keywords selected, there is a higher chance that two or more researchers will select the same keyword by chance.[5] Consequently, since journals with a higher number of keywords (K_j) tend to show a higher number of occurrences (OCC_j), these journals may show a higher value of the specialization index *Sj*. In other words, if we calculate the *Sj* of two journals with a very different number of keywords, the journal with the higher number of keywords tends to be overestimated.

The positive linkage between K_j and OCC_j then makes a correction of *Sj* necessary for those cases where K_j is higher to avoid overestimating *Sj*.

Thus, we propose an adjusted version of *Sj* that can be written as follows:

$$S_{J-adj} = \frac{\ln(OCC_j)}{\ln(K_j)} = \log_{K_J}(OCC_j) \qquad (2)$$

Using the logarithm of variables represents an attempt to reduce the impact of outliers and extreme values that may skew the analysis results. While the correction proposed by [2] should mitigate the bias of the *K-OCC* linkage, it has the drawback of losing the immediate meaning of *Sj* (i.e., the number of times a keyword is on average repeated in a journal) and is only one of the possible corrections. In reality, one could also skip the use of such a measure of correction if the *Sj* calculation is limited to a set of journals with approximately the same number of keywords.

---

[4] Since the number of keyword is a subset of the set of words of a language, ideally W is infinite because language users constantly coin new words (Kornai, 2002).

[5] Let's take a contrary example. Suppose a journal has only one article and one keyword. Now suppose the journal accepts a second article with a second keyword. The probability that this second keyword is equal to the first is very low, considering the potential number of keywords. However, if an increasing number of articles (and keywords) is accepted, the probability of having a keyword repeated more than once increases.

The dataset

The data used to calculate the journal specialization measure proposed in this study (*Sj*) were based on keyword frequencies obtained from 50 journals selected from the Scopus source title list (updated in March 2023). More specifically, we selected 50 active journals from the Scopus journal list in alphabetical order containing a minimum number of unique keywords[6]. For each journal, we retrieved all their articles from the Scopus database. The maximum threshold of 20,000 journal articles imposed in the Scopus algorithm has not represented a problem since no one of the 50 journals has more than 20,000 documents to export. Each journal is classified according to the Scopus All Science Journal Classification, A*SJC* (Franceschini et al., 2016). The keywords were collected through the function available in VOSviewer (version 1.6.18), a freely available bibliometric software explicitly developed to create and display bibliometric maps (Eck & Waltman, 2010, 2014). This function permits the collection of unique keywords and their occurrences at the article level. For keywords, this study intends only author keywords, the terms selected by authors to accurately represent the content of their document. The indexed keywords—that represent keywords selected and standardized by Scopus for indexing purposes using vocabularies derived from Elsevier-owned or licenced thesauri—are excluded from the analysis even though they represent a separate vector of data that deserves further investigation in future research.[7]

The number of keywords and occurrences of each journal and some additional details are reported in Table 1. We collected 119,775 unique keywords[8] in total. At a single journal level, the number of keywords varied from a minimum of 43 (for the journal *AAO Journal*) to a maximum of 13,112 (for the journal *Accident Analysis and Prevention*).

Using a data cleaning function able to merge different keyword variants in unique forms (a VOSviewer thesaurus file), the keywords collected in this study were submitted to a filter to delete too generic or ambiguous keywords that do not consent to specify in detail the content of an article. The reasons for the ambiguousness vary from the extreme generalization of keywords (genetics, health, chemistry) to an incorrect specification of the article content (e.g., case, report, note, etc.) up to confounding terms (e.g., united states, adult, procedures).[9] In making this correction, it is necessary to clarify that it is not meant to label these keywords as inappropriate. Rather, the purpose of the correction is to highlight

---

[6] The minimum number has been arbitrarily set at 40. It should be noted that not all journals have retrievable keywords. In some cases, there may be a lack of author keywords, index keywords, or both. Since the topic of missing keywords remains unexplored, it could be a worthwhile area for future research.

[7] More specifically, index keywords are keywords not entered by authors but from the reference database (e. g. Scopus or WoS). Scopus, for example, has index keywords that use a controlled vocabulary to describe the contents of a study, such as MeSH (medical subject headings), Emtree (life sciences & health science), or Compendex (engineering).

[8] In this study the term "unique" is referred to the single journal *j* and not to the set of 50 journals since it is not possible to exclude the possibility that a unique keyword in one journal also appears in another journal. Table 1 Keywords and occurrences of Journals analysed.

[9] The list of the keywords qualified as ambiguous and then deleted from our data is the following: covid-19, Covid19, COVID/19, covid_19, adult, aged, animal, animals, article, case report, chemistry, clinical article, clinical trial, conference paper, controlled study, diagnosis, diseases, editorial, female, genetics, health, human, humans,letter, major clinical study, male, methodology, nonhuman, note, priority journal, procedures, review, short survey, therapy, united kingdom, united states, europe. We also removed the keyword "covid-19" or similar terms (i.e. Covid19, COVID/19, covid_19) due to the massive impact of the COVID-19 pandemic on scientific production (Riccaboni and Verginer, 2022) and the related possible bias in keyword distribution.

Table 1 Keywords and occurrences of Journals analysed

(1) Journals$_{(j)}$	(2) NP$_j$	(3) K$_j$	(4) OCC$_j$	(5) NC$_j$	(6) Time span
2D Materials	1728	4115	7839	7.42	2014:2023
3 Biotech	3297	10,310	16,693	5.14	2012:2023
3D Printing and Additive Manufacturing	340	842	1210	4.58	2014:2023
3 L: Language, Linguistics, Literature	594	2281	2891	2.16	2008:2022
452°F	123	413	471	1.09	2019:2022
4OR	513	1465	2066	2.34	2003:2023
A e C—Revista de Direito Administrativo e Constitucional	111	435	545	1.72	2019:2022
a/b: Auto Biography Studies	408	320	380	4.52	2015:2022
AAC: Augmentative and Alternative Communication	892	1435	3304	2.75	1985:2023
AACE Clinical Case Reports	695	476	627	4.28	2015:2023
AACN Advanced Critical Care	868	1190	1816	2.07	2006:2023
AANA Journal	2691	3092	4885	2.08	1973:2022
AAO Journal	111	43	43	1.78	2009:2020
AAPS Advances in the Pharmaceutical Sciences Series	689	2249	3197	2.8	2011:2022
AAPS Journal	2076	5692	9612	5.56	1999:2023
AAPS PharmSciTech	266	868	1004	3.56	1999:2006
AATCC Journal of Research	326	1222	1772	3.78	2014:2023
AATCC Review	957	875	1558	2.01	2001:2020
ABAC Journal	223	816	910	2.15	2018:2023
Abacus	1060	1707	2782	1.81	1965:2023
Abdominal Radiology	3312	5384	13,030	6.25	2016:2023
Abhandlungen aus dem Mathematischen Seminar der Universitat Hamburg	1850	1080	1223	1.26	1922:2023
Abstract and Applied Analysis	5204	108	114	2.5	1996:2023
Academia (Greece)	170	553	734	1.78	2017:2023
Academia Revista Latinoamericana de Administracion	274	992	1283	2.82	2013:2023
Academic Emergency Medicine	6658	5605	15,208	4.61	1994:2023
Academic Forensic Pathology	712	1281	2776	2.24	2011:2022
Academic Journal of Manufacturing Engineering	864	2777	3574	2.67	2010:2022
Academic Journal of Second Military Medical University	4956	10,326	19,170	4.94	2006:2022
Academic Medicine	14,803	850	3562	3.1	1940:2023
Academic Pathology	371	1333	2346	4.73	2014:2023
Academic PEDIATRICS	1996	3054	7187	5.44	2010:2023
Academic Psychiatry	1936	1249	2712	3.78	1989:2023
Academic Questions	1604	102	109	1.1	1988:2020
Academic Radiology	7144	9511	21,518	5.14	1994:2023
Acarina	267	666	1549	2.2	2009:2022
Acarologia	671	1803	3568	2.96	1977:2023
Access to Justice in Eastern Europe	186	761	973	1.88	2018:2023
Accident Analysis and Prevention	7534	13,112	31,139	3.38	1969:2023
Accountability in Research	815	1759	3298	2.38	1989:2023
Accounting and Business Research	1692	1445	2033	1.68	1970:2023
Accounting and Finance	1734	3893	6266	2.51	1979:2023
Accounting Education	1008	1903	3058	1.98	1992:2023

Table 1 (continued)

(1) Journals$_{(j)}$	(2) NP$_j$	(3) K$_j$	(4) OCC$_j$	(5) NC$_j$	(6) Time span
Accounting Forum	470	1534	2143	2.37	2004:2023
Accounting Historians Journal	684	301	337	1.52	1974:2022
Accounting History	610	1592	2476	1.8	1996:2023
Accounting History Review	156	514	657	1.69	2011:2022
Accounting Horizons	914	1506	2191	2.82	1996:2023
Accounting Research Journal	427	1316	1903	2.4	2005:2023
Accounting Review	1593	3619	6713	2.42	1996:2023

This table reports the number of keywords (variable K_j, column 3) retrieved from Scopus for the 50 journals listed in column 1. The number of articles collected for each journal is indicated in column 2 by *NP*. The variable *OCC* (column 4) is the total number of occurrences of keywords appearing in the journal. Variable *NC* reported in column 5 indicates the average number of coauthors per document. In column 6 are reported the time-span of the articles collected for each journal. Data have been retrieved in Scopus at the date of May 1st 2023

the importance of correcting keywords during estimation exercises to prevent overestimation of the specialization index.

Through the function provided by Bibliometrix (Aria & Cuccurullo, 2017), we also counted the average number of coauthors per document for each journal (see *NC* in Table 1). The number of papers retrieved on Scopus (NP$_j$) has been calculated with Bibliometrix and was not the same for all journals and varied from 111 (the number of papers for *A e C—Revista de Direito Administrativo e Constitucional* and for *AAO Journal*) to 14,803 (for *Academic Medicine)*. The period of collection also differed, as the briefest time span was 4 years, from 2019 to 2022 (for *A e C—Revista de Direito Administrativo e Constitucional*), and the longest was 1922–2023 (for *Abhandlungen aus dem Mathematischen Seminar der Universitat Hamburg*).

Keyword dynamics across journals

This section analyses the dynamics of the number of occurrences of keywords in a journal. At first glance, the occurrence of keywords (OCC_{ij}, the number of times the i_{th} keyword appears in the journal j_{th}) is a function of the number of keywords. Specifically, by using the analogy of a deck of cards (i.e., an actual finite population from which samples can be randomly drawn) and considering that the topics of articles published in a volume of a journal should be random, we can approximate the choice of a keyword as a card taken from a deck of W_j cards with replacement (where W_j is the unknown and many possible keywords), while the occurrence should be approximated as success in selecting a keyword already extracted by another author. Under these circumstances, the number of occurrences is increasing in the number of keywords, as the drawing with replacement analogy is positively associated with the number of trials.

The positive linkage between the number of occurrences and the number of keywords is empirically tested in Fig. 1. The special connection between keywords and occurrences is observed in all journals, independent of the A*SJC* macrodiscipline.
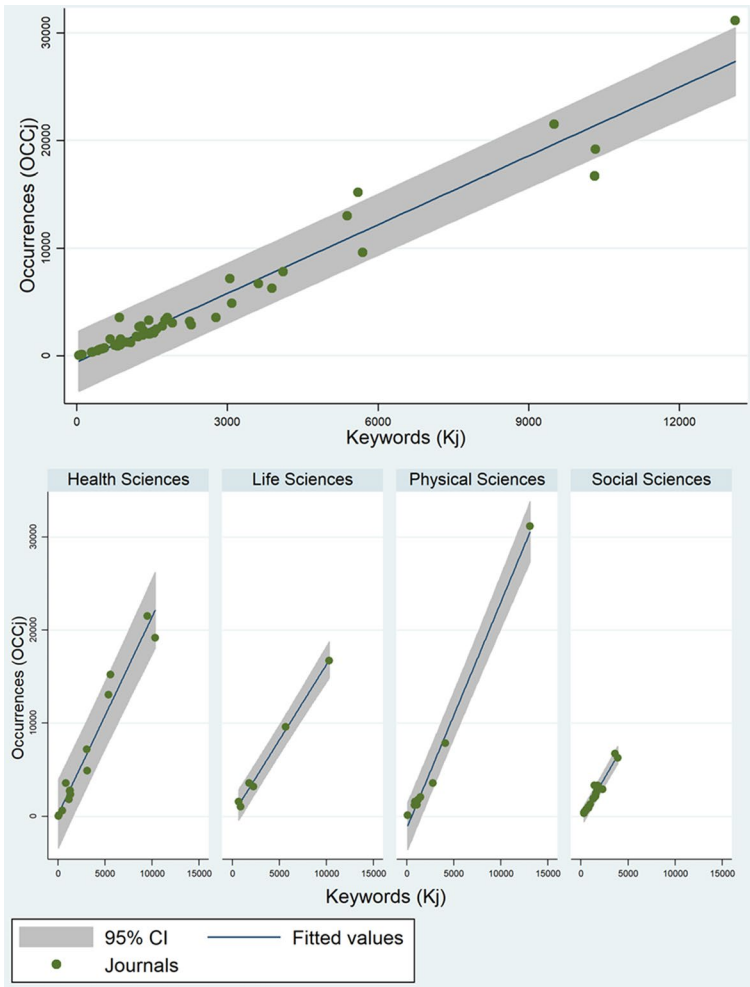
This figure displays the connection between the number of unique keywords and their occurrences among the 50 journals constituting the sample: subpicture a) reports the case of all journals; subpicture b) groups journals according to the macrodiscipline categorization reported by the Scopus All Science Journal Classification (ASJC). The grey area indicates the 95% confidence interval for the predicted values.

Fig. 1 Keyword and occurrences relationship

Since the number of keywords seems, at this point, to reasonably act as a factor that affects the number of occurrences, one wonders what are the main factors that determine the number of keywords appearing in a set of papers. The first possible answer is to connect the number of keywords to the size of the set of papers analysed since one should expect that a high number of papers analysed corresponds to a high number of topics investigated and hence a high number of keywords. As shown in Fig. 2a, this relationship is observed in our sample.

However, analysing the dynamics of keywords is more complex, and it seems to represent a still underexplored subject. For example, the keyword dynamics should be a function not only
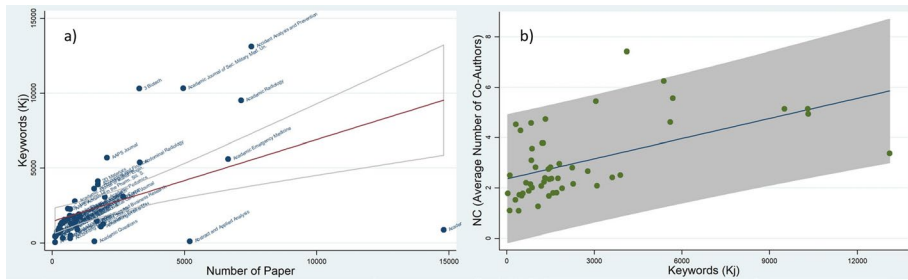
Subpicture a) reports the scatter plot of the connection between the number of unique keywords and the average number of papers collected. The long-dashed line indicates the area's border containing the 95% confidence interval for the predicted values. *Subpicture b)* reports the scatter plot between the number of unique keywords and the average value of the number of coauthors per document; each circle represents a journal, while the grey area indicates the 95% confidence interval for the predicted values.

Fig. 2 Keywords, number of papers, and number of coauthors

of the number of collected papers but also of the journal submission settings to the extent that some journals may require a minimum number and/or a maximum number of keywords.

Additional factors that could explain a high number of keywords for a single paper could be the number of article pages, as the number of pages could reasonably be correlated with the article's depth (and then to the number of keywords selected). Not being able to directly correlate the number of pages with the number of keywords for each article, this study considered the average number of authors for each journal since a relationship between article length and the number of authors was found, albeit weak and limited to a single discipline (Papatheodorou et al., 2008). The image shown in Fig. 2b supports this idea and suggests considering this relationship in future research related to inspecting keyword dynamics across academic journals.

Results

This section shows the results of calculating the level of concentration of keywords in a set of articles of journal j, with the Sj index exhibited in [1] that we use as a proxy of journal specialization.

Recalling that the Sj index indicates the number of times a keyword appears on average in the j_{th} journal, the value of the Sj index for each of the 50 journals included in the sample is shown in Table 2. The highest level of Sj is 4,19, which is attributed to *Academic Medicine,* while the lowest level ($Sj = 1$) is related to *AAO Journal*.

Since a pattern of dependence of keywords on the number of papers was observed and discussed in the previous section, the journals with the higher Sj are not surprisingly almost coincident with the journals with the higher number of papers collected.

In particular, if we qualify as high-frequency journals (*hfj*) the journals with numerous collected papers higher than the 75_{th} percentile of NP_j (equal to 1850), we observe that 7 out of 10 journals with the higher Sj are in fact *hfj*. Moreover, the correlation between Sj and NP_j is 0.74 (highly statistically significant at the 1% level), indicating a positive relationship that can be graphically observed in Fig. 3a.

The positive relationship between the number of papers and the specialization index also emerges if we observe the dynamic of Sj inside the same journal. Specifically, if we analyse the Sj of a single journal, we observe a tendency of the Sj index of that

Table 2 Journal ranking by specialization

(1) Journal	(2) size	(3) Sj	(4) Rank Sj	(5) Sj-adj	(6) Rank Sj-adj	(7) Δ Rank
2D Materials	(*lfj*)	1.9050	12	1.0774	15	− 3
3 Biotech	(*lfj*)	1.6191	20	1.0521	27	− 7
3D Printing and Additive Manufacturing	(*lfj*)	1.4371	29	1.0538	25	+ 4
3 L: Language. Linguistics. Literature	(*lfj*)	1.2674	40	1.0306	41	− 1
452°F	(*lfj*)	1.1404	44	1.0218	43	+ 1
4OR	(*lfj*)	1.4102	31	1.0472	30	+ 1
A e C—Revista de Direito Administrativo e Constitucional	(*lfj*)	1.2529	41	1.0371	38	+ 3
a/b: Auto Biography Studies	(*lfj*)	1.1875	42	1.0298	42	*No change*
AAC: Augmentative and Alternative Communication	(*lfj*)	2.3024	7	1.1147	4	+ 3
AACE Clinical Case Reports	(*lfj*)	1.3172	35	1.0447	35	*no change*
AACN Advanced Critical Care	(*lfj*)	1.5261	25	1.0597	22	+ 3
AANA Journal	(hfj)	1.5799	23	1.0569	24	− 1
AAO Journal	(*lfj*)	1.0000	50	1.0000	50	*No change*
AAPS Advances in the Pharmaceutical Sciences Series	(*lfj*)	1.4215	30	1.0456	33	− 3
AAPS Journal	(hfj)	1.6887	18	1.0606	20	− 2
AAPS PharmSciTech	(*lfj*)	1.1567	43	1.0215	44	− 1
AATCC Journal of Research	(*lfj*)	1.4501	27	1.0523	26	+ 1
AATCC Review	(*lfj*)	1.7806	16	1.0852	12	+ 4
ABAC Journal	(*lfj*)	1.1152	47	1.0163	47	*No change*
Abacus	(*lfj*)	1.6298	19	1.0656	18	+ 1
Abdominal Radiology	(hfj)	2.4201	3	1.1029	8	− 5
Abhandlungen aus dem Mathematischen Seminar der Universitat Hamburg	(hfj)	1.1324	45	1.0178	46	− 1
Abstract and Applied Analysis	(hfj)	1.0556	49	1.0115	49	*No change*
Academia (Greece)	(*lfj*)	1.3273	34	1.0448	34	*No change*
Academia Revista Latinoamericana de Administracion	(*lfj*)	1.2933	36	1.0373	37	− 1
Academic Emergency Medicine	(hfj)	2.7133	2	1.1156	3	− 1
Academic Forensic Pathology	(*lfj*)	2.1671	10	1.1081	6	+ 4
Academic Journal of Manufacturing Engineering	(*lfj*)	1.2870	37	1.0318	40	− 3
Academic Journal of Second Military Medical University	(hfj)	1.8565	14	1.0669	17	− 3
Academic Medicine	(hfj)	4.1906	1	1.2124	1	*No change*
Academic Pathology	(*lfj*)	1.7599	17	1.0786	14	+ 3
Academic Pediatrics	(hfj)	2.3533	5	1.1067	7	− 2
Academic Psychiatry	(hfj)	2.1713	9	1.1087	5	+ 4
Academic Questions	(*lfj*)	1.0686	48	1.0144	48	*No change*
Academic Radiology	(h)	2.2624	8	1.0891	11	− 3
Acarina	(*lfj*)	2.3258	6	1.1298	2	+ 4
Acarologia	(*lfj*)	1.9789	11	1.0910	10	+ 1
Access to Justice in Eastern Europe	(*lfj*)	1.2786	38	1.0370	39	− 1

Table 2 (continued)

(1) Journal	(2) size	(3) Sj	(4) Rank Sj	(5) Sj-adj	(6) Rank Sj-adj	(7) Δ Rank
Accident Analysis and Prevention	(hfj)	2.3748	4	1.0912	9	− 5
Accountability in Research	(lfj)	1.8749	13	1.0841	13	No change
Accounting and Business Research	(lfj)	1.4069	32	1.0469	31	+ 1
Accounting and Finance	(lfj)	1.6096	21	1.0576	23	− 2
Accounting Education	(lfj)	1.6069	22	1.0628	19	+ 3
Accounting Forum	(lfj)	1.3970	33	1.0456	32	+ 1
Accounting Historians Journal	(lfj)	1.1196	46	1.0198	45	+ 1
Accounting History	(lfj)	1.5553	24	1.0599	21	+ 3
Accounting History Review	(lfj)	1.2782	39	1.0393	36	+ 3
Accounting Horizons	(lfj)	1.4548	26	1.0512	29	− 3
Accounting Research Journal	(lfj)	1.4460	28	1.0514	28	No change
Accounting Review	(lfj)	1.8549	15	1.0754	16	− 1

Sj the number of times that a keyword on average appears in the j_{th} journal. *Sj-adj* is a version of *Sj* corrected for the number of keywords. Rank *Sj* and Rank *Sj-adj* represent the position of a journal in the ranking order for *Sj* and *Sj-adj*, respectively. Δ Rank is the difference between Rank *Sj* and Rank *Sj-adj*. Column (2) marks high-frequency journals (hfj) and low-frequency journals *(lfj)*: journals indicated with (hfj) have numerous articles equal or higher than 1850 (the 75th percentile of column 2 of Table 1); journals indicated with *(lfj)* have numerous articles less than 1850

journal to be positively correlated with the number of papers of the sample from which to collect the number of keywords. Figure 3b reports the case of the two journals *2D Materials* and *3 Biotech* (the first and the second journals of the sample in alphabetical order), and it is possible to observe that *Sj* is also positively correlated with the number of papers in this case.

The positive relationship between the index of specialization of a journal and the number of articles retrieved to collect the keywords raises concerns about the appropriateness of calculating the *Sj* index among journals with different sizes of article collection.

One possible solution could be to limit the comparison of any keyword-based metric, such as *Sj*, only to journals for which the same amount of research output (i.e., NP_j) was collected.

An alternative approach could be to normalize the *Sj* index for the natural logarithm of K_j, as shown by equation [2].

The adjusted version of the keyword density level, *Sj-adj* and the relative rank were then calculated for each journal, and the results are presented in columns 5 and 6 of Table 2, respectively.

The adoption of the correction changes the rank of 40 out of 50 journals, while the other 10 (20,0% of the sample) remain unaltered, including the journal positioned in the top rank (i.e., *Academic Medicine*). However, the amplitude of change in ranks after the adoption of the *Sj-adj* measure is minimal. Specifically, 64% of journals show a change in the ranks of only ± 3 ranks, while approximately 16% of journals show a change higher than 3 notches with respect to the *Sj*-based ranking. A large change is observed for *3 Biotech*, which scaled down the ranking by 7 positions. Taken together, these findings suggest that using
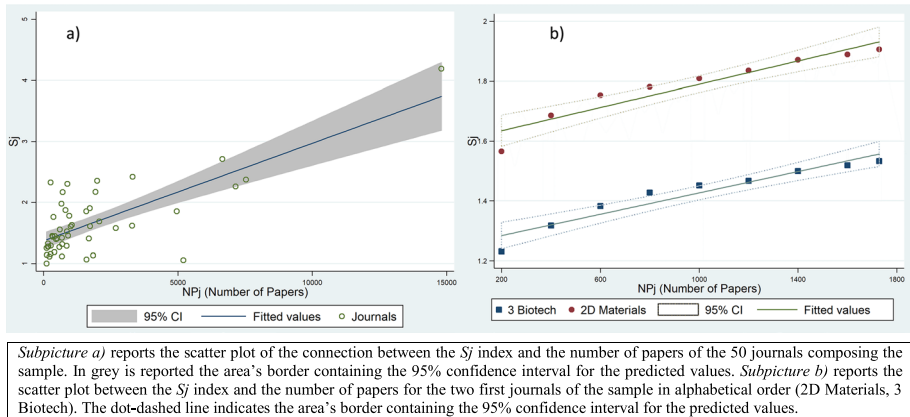
Subpicture a) reports the scatter plot of the connection between the *Sj* index and the number of papers of the 50 journals composing the sample. In grey is reported the area's border containing the 95% confidence interval for the predicted values. *Subpicture b)* reports the scatter plot between the *Sj* index and the number of papers for the two first journals of the sample in alphabetical order (2D Materials, 3 Biotech). The dot-dashed line indicates the area's border containing the 95% confidence interval for the predicted values.

Fig. 3 Index of specialization and number of papers retrieved

Sj-adj does not radically alter the previous *Sj*-based results, as is also revealed by Fig. 4, which compares the rank of each journal according to the two measures discussed above.

Conclusions and limitation

The number of academic journal titles has grown progressively during the last century, and the scholarly publication system is experiencing new growth because of Internet technology and open access options (Gu & Blackmore, 2016). This study introduces a new methodology for constructing an index of journal specialization. Based on the collection of keywords appearing in a predefined collection of articles published by a journal, the study proposes to measure journal specialization through the number of occurrences of keywords appearing in a journal. In this section, we discuss the strengths and limitations of the proposed methodology. Important strengths are the extreme simplicity of the methodology and its immediate comprehensibility, as the *Sj* index could be interpreted as the number of times that a keyword appears on average among the articles published by a specific journal. The higher this number is, the higher the fraction of papers about a specific research subject. The underlying assumption is that the keywords are chosen carefully by researchers to adequately represent the content of manuscripts and to be specific to a field or subfield of research.

The methodology is fully documented in this paper and accompanied by an analysis of keyword dynamics that, to our knowledge, has not been explored before. To determine journal specialization, the methodology relies exclusively on unique keywords and occurrences that can be easily retrieved through VOSviewer (Eck & Waltman, 2010). The data observed for our sample of 50 journals suggest considering the number of papers (NP_j) collected for the j_{th} journal as a crucial variable since the number of keywords K_j is dependent on the amount of research analysed. Furthermore, K_j positively affects the occurrence of OCC_j, which affects the same *Sj* index (since it is calculated as OCC_j/K_j).

To avoid any overestimation of the *Sj* index, the study proposes an adjustment of *Sj* for the number of keywords; the corrected version of the index seems not to be changed overall to the extent that more than 80% of the papers collected do not show a significant variation in specialization ranks.
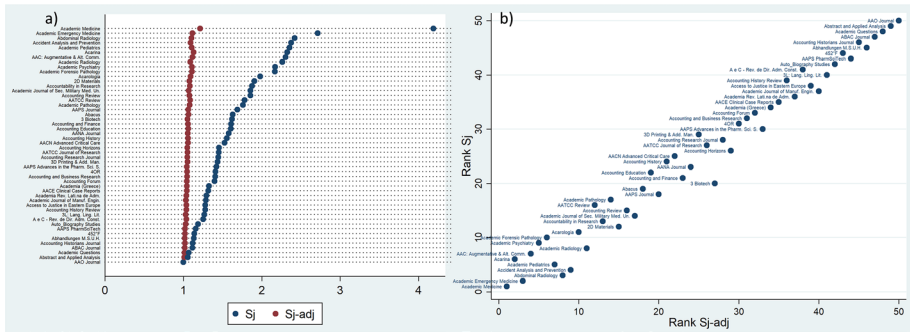
This figure reports in graphic form the values reported in Table 2. Note: a) the left dot graph indicates the dot graph of *Sj* and *Sj*-adj values, while the right scatter plot shows the correlation between the ranks obtained with *Sj* and *Sj*-adj for each journal.

Fig. 4 This figure reports in graphic form the values reported in Table 2. Note: the left dot graph reported in subpicture *a*) indicates the dot graph of S*j* and S*j-adj* values, while the right scatter plot reported in subpicture *b*), shows the correlation between the ranks obtained with S*j* and S*j-adj* for each journal

Nevertheless, the findings of this study must be seen in light of some limitations. First, the *Sj* index cannot definitively conclude without caution the specialization of a journal in a specific discipline. The index only indicates the level of overlapping articles' contents, assuming that they are reasonably approximated by the keywords. From an abstract perspective, a journal could show a high level of *Sj*, while its articles may cover different disciplines. In a basic example regarding the recent or structural events, a journal could have published many papers with the keyword "COVID-19" or "climate" even though the contents cover various research fields (medicine, economics, engineering, sport science, etc.).

A second limitation is that the *Sj* index could be extremely affected by an underestimation of the number of occurrences OCC_j, as the matching of keywords could fail due to the different ways in which two authors can write the same keyword. The number of examples is indefinite, as the discoupling of two similar keywords may occur due to recourse to the use of singular or plural, or different uses of special characters (e.g., *fibrillin 1* rather than *fibrillin-1*), or synonyms.[10] Third, to conclude about the reliability of the *Sj* index, one should consider how it is affected by many aspects related to the size and the quality of the keywords retrieved that may depend on many factors, such as (i) the attitude of the researchers towards inserting detailed article-specific keywords, (ii) the choice of the journal to impose a minimum (and/or a maximum) number of keywords per article (Golub et al., 2020) and (iii) the lack of a standard for the length of keywords.[11]

The ranking's high sensitivity to multiple factors highlights that the purpose of this work is not to determine the ranking of specialized journals but rather to contribute to the ongoing discussion on specialization and interdisciplinarity metrics. Importantly, the sample size used in this study is arbitrary and limited to only 50 journals out of a total of approximately 44,000 journals listed in Scopus as of March 2023 (active and inactive journals).

---

[10] For example, the keyword interest rate in the banking discipline may have at least eight synonyms: annual percentage rate, bank rate, borrowing rate, lending rate, price of money, prime interest rate, and prime rate.

[11] In fact the most common practice is not inserting indications on the length of the keywords, although in some journals (also journal included in our sample) it is reported that: "each individual keyword may be up to 256 characters in length".

This study could also help address future research that intends to use keywords (and lemmatization of keywords) to enhance the journal specialization debate. For example, an issue that needs to be addressed in future works is the introduction of measures that also consider the characterization of the distribution of keywords (e.g., skewness and kurtosis). In this regard, there is room for improvement in the approach proposed here by adapting metrics such as the Herfindahl–Hirschman index (HHI) to the use of keywords and occurrences.

References

Abramo, G., D'Angelo, C. A., & Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics, 12*(4), 1182–1193.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975.

Augsburg, T. (2016). *Becoming Interdisciplinary: An Introduction to Interdisciplinary Studies*. Kendall Hunt.

Bensman, S. J. (2001). Bradford's law and fuzzy sets: Statistical implications for library analyses. *IFLA Journal, 27*, 238–246.

Boyack, K. W., & Klavans, R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines. Paper Presented at the 13th Conference of the International Society for Scientometrics and Informetrics, Durban

Boyack, K. W., & Klavans, R. (2010). Co-Citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389–2404.

Braam, R. R., Moed, H. F., & Van Rann, A. F. J. (1991). Mapping of science by combined co-citation and word analysis I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Callon, M., Law, J., & Rip, A. (1986). Qualitative scientometrics. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world.* The Macmillan Press Ltd.

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

Choi, J., Yi, S., & Lee, K. C. (2011). Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management, 48*(8), 371–381.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics, 10*, 933–953.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/ NC-value method. *International Journal on Digital Libraries, 3*(2), 115–130.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national. *Regional and Institutional Level Scientometrics, 37*(2), 195–221.

Glänzel, W., & Debackere, K. (2022). Various aspects of interdisciplinarity in research and how to quantify and measure those. *Scientometrics, 127*, 5551–5569.

Glänzel, W., Schubert, A., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics, 44*, 427–439.

Golub, K., Tyrkkö, J., Hansson, J., & Ahlström, I. (2020). Subject indexing in humanities: A comparison between a local university repository and an international bibliographic service. *Journal of Documentation, 76*(6), 1193–1214.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(suppl 1), 5228–5235.

Gu, X., & Blackmore, K. (2016). Recent trends in academic journal growth. *Scientometrics, 108*, 693–716.

Hartley, J., & Kostoff, R. N. (2003). How useful are `key words' in scientific journals? *Journal of Information Science, 29*(5), 433–438.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133–159.

Jarneving, B. (2001). The cognitive structure of current cardiovascular research. *Scientometrics, 50*(3), 365–389.

Kornai, A. (2002). How many words are there? *Glottometrics, 4*, 61–86.

Leydesdorff, L. (2007). Mapping interdisciplinarity at the interfaces between the science citation index and the social science citation index. *Scientometrics, 71*, 391–405.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87–100.

Li, M. (2018). Classifying and ranking topic terms based on a novel approach: Role differentiation of author keywords. *Scientometrics, 116*(1), 77–100.

Milojević, S. (2020). Practical method to reclassify web of science articles into unique subject categories and broad disciplines. *Quantitative Science Studies, 1*(1), 183–206.

Onyancha, O. B. (2018). Forty-five years of LIS research evolution, 1971–2015: An informetrics study of the author-supplied keywords. *Publishing Research Quarterly, 34*(3), 456–470.

Papatheodorou, S. I., Trikalinos, T. A., & Ioannidis, J. P. A. (2008). Inflated numbers of authors over time have not been just due to increasing research complexity. *Journal of Clinical Epidemiology, 61*(6), 546–551.

Riccaboni, M., & Verginer, L. (2022). The impact of the COVID-19 pandemic on scientific research in the life sciences. *PLoS One*. https://doi.org/10.1371/journal.pone.0263001

Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics, 13*(1), 202–225.

Shu, F., Ma, Y., Qiu, J., et al. (2020). Classifications of science and their effects on bibliometric evaluations. *Scientometrics, 125*(3), 2727–2744.

Tsai, L. C., Hwang, S. L., & Tang, K. H. (2011). Analysis of keyword-based tagging behaviors of experts and novices. *Online Information Review, 35*, 272–290.

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics, 10*, 1166–1177.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*, 523–538.

van Eck, N. J., & Waltman, L. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*, 2378–2392.

van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 285–320). Springer.

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies, 1*(1), 239–263.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science, 19*, 473–496.

Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. *Scientometrics, 117*, 973–995.

Zwanenburg, S., Nakhoda, M., & Whigham, P. (2022). Toward greater consistency and validity in measuring interdisciplinarity: A systematic and conceptual evaluation. *Scientometrics, 127*, 7769–7788.