Identifying the Main Research Topics in Open Access Journals: An Analysis with Bibliometric Metrics

Patrícia M. Dias[1], Thiago M. R. Dias[2], Gray F. Moita[2]

[1]*State University of Minas Gerais*
Av. Paraná, 3001 - Jardim Belvedere – Zip-Code: 35501-170 - Divinópolis – MG - Brazil.
patriciamdias@gmail.com
[2]*Federal Center for Technological Education of Minas Gerais*
Rua Álvares de Azevedo 400 - Bela Vista – Zip-Code: 35503-822 - Divinópolis - MG - Brazil
thiagomagela@cefetmg.br, gray@cefetmg.br

Abstract. The traditional printed format of science communication is gradually giving way to new electronic formats, due to the rise of information and communication technology. In the context of research and scientific studies, scientific communication appears today as a central element at different levels of discussion, with emphasis on the dissemination of scientific articles in journals, currently one of the main means of communication for this purpose. In the context of this work, in order to better understand the main research topics investigated by Brazilian researchers in open access journals, the curriculum data repository of the Lattes Platform was used. Currently, the Lattes Platform has a set of more than 7 million registered CVs.

Keywords: Research Topics, Scientific Production, Lattes Platform..

1 Introduction

The traditional printed format of scientific communication has gradually given way to new electronic formats, due to the rise of information and communication technology. In the context of scientific research and studies, scientific communication has emerged today as a central element at various levels of discussion, with an emphasis on the dissemination of scientific articles in journals, currently one of the main means of communication for this purpose.

Mueller [1] states that scientific journals perform at least four essential functions: certification of science with the endorsement of the scientific community; communication channel between scientists and broader dissemination of science; scientific archive; and record of the authorship of scientific discovery.

In this context, open access scientific publishing is part of a broader scenario in favor of the opening of knowledge in general (open access, open data, open educational resources, free software, open licenses) and essentially constitutes a movement towards the conception of information and knowledge as public goods [2].

Given that a large part of scientific research in the country is funded with public resources, usually in public educational institutions or research centers, it is expected that the results of such studies will be published without any type of barrier, especially financial ones. In this context, combined with the advantages that open access publications offer, such as availability, visibility and accessibility, several efforts are being made to ensure that more and more scientific articles are published in open access journals.

In view of this, understanding which main research topics are being investigated in articles published in open access journals makes it possible to identify an overview of the main themes studied. It also allows us to verify the representativeness of certain topics present in the articles analyzed.

2 Methodology

Bibliometrics aims to develop standards and mathematically model processes for measurements and, based on the results, draw predictions and make possible decisions.

Through its techniques, bibliometrics seeks to study the quantitative aspects of science and scientific production as an activity that involves social, economic and political characteristics. It provides a tool for studies that aim to map scientific knowledge and extract information, as well as understanding how scientific production has been carried out [3].

Among the main bibliometric laws is Zipf's Law. Zipf's Law is related to the frequency of occurrence of words in a given text. This law developed and extended an empirical law observed by Estoup in 1916, which establishes a relationship between the position of a word and the frequency of its appearance in a long text. Zipf's Law is formulated as follows: r.f = c, where "r" is the position of the word, "f" is the frequency and "c" is a constant. Zipf derived his law from a general principle of "minimum effort," according to which a word whose usage cost is low or whose transmission demands minimal effort is frequently used in a large text [4].

In the context of this work, in order to better understand the main research topics investigated by Brazilian researchers in open access journals, the curriculum data repository of the Lattes Platform was used. The Lattes Platform currently has a set of more than 7 million registered curriculum vitae.

To define the data set to be analyzed in this section, it was decided to extract the words from the titles of articles published in open access journals (2,090,015). The choice to extract the words from the titles of the articles instead of the keywords linked to the articles is due to the fact that approximately only 17% of the articles analyzed had keywords linked to them. In addition, several studies have used the words in the titles of publications as the object of analysis [5,6,7,8].

In addition, considering that the registration of keywords of a scientific article in their CVs is the sole responsibility of the respective researchers, and this is done freely by them, it means that any set of characters can be inserted as a keyword. From this, there is usually a very large collection of keywords without any pattern [9].

Therefore, for the analyses performed here, the titles of all publications in the identified set were considered. The titles underwent a data processing process that aimed to identify the words that will later be the object of analysis. All stages of the processing process can be seen in Table 1.

Table 1 – Stages of the Data Processing Process.

Algorithm Steps	Results
Recebimento do Título	UMA ESTRATÉGIA PARA IDENTIFICAÇÃO DE ARTIGOS EM PERIÓDICOS DE ACESSO ABERTO NA PLATAFORMA LATTES.
LowerCase	uma estratégia para identificação de artigos em periódicos de acesso aberto na plataforma lattes
StopWords_PT	estratégia identificação artigos periódicos acesso aberto plataforma lattes
StopWords_EN	estratégia identificação artigos periódicos acesso aberto plataforma lattes
Identificação de Termos	estratégia
	identificação
	artigos
	periódicos
	acesso
	aberto
	plataforma
	lattes

As can be seen, each article's title is retrieved and, in this way, the data processing process is initialized. In the LowerCase step, all words are converted to lowercase in order to standardize the set, as well as to prevent words from being mapped to different topics because some have uppercase letters and others do not. In the stopWords removal process (StopWords_PT and StopWords_EN), all terms that do not have significant semantic values to characterize a research topic are removed, thus reducing the volume of words to be processed and analyzed. StopWords were initially removed in Portuguese and later in English, given that these are the most widely used languages, as already presented.

Since the initial object of analysis is the title of the articles, in which there is a concern with the general description of the study to be presented, the number of stopWords is significant, unlike the keywords, justifying their removal for the analyses to be performed. Afterwards, in the last stage, Term Identification, the words are separated into topics, which will make up a dictionary of terms for counting frequencies.

3 Results

Initially, 28,636,958 topics were identified, considering all the words in the article titles. After removing duplicates, the set was reduced to 423,364 unique words. Subsequently, with the removal of stopWords, the set had a total of 393,896 words that became the object of analysis. Table 2 presents the terms with the highest frequency before the removal of stopWords.

Table 2 – Frequency of Main Terms without Removal of StopWords

Position	Frequency	Word
1	1.741.809	de
2	762.560	of
3	716.058	e
4	529.940	a
5	513.477	do
6	507.828	da
7	501.705	in
8	463.628	and
9	452.657	em
10	373.515	the
11	234.604	no
12	230.966	na
13	217.452	o
14	156.366	para
15	149.550	com

As can be seen, and considering the characteristics of the titles of publications that generally require the use of stopWords in their composition, all of the first 15 terms identified are stopWords in Portuguese or English. Given the very significant frequency of these terms, their exclusion was justified.

Approximately 64% of the publications in open access journals analyzed in this study are in Portuguese, which justifies the considerable number of terms in this language. Table 3 presents the result of the extraction and ordering by frequency of the words in the titles of each article analyzed, after all the data processing.

Table 3 – Distribution of words by position (x) and their frequencies (y)

Position (x)	Frequency(y)	Word
1	88.485	brazil
2	85.470	brasil
3	72.100	estudo
4	71.618	avaliação
5	69.314	análise
6	58.977	saúde
7	46.863	educação
8	44.131	study
9	43.411	brazilian
10	42.365	rio
11	41.411	patients
12	38.754	diferentes
13	37.788	produção
14	37.586	ensino
15	37.395	estado
⋰	⋰	⋰
393.894	1	zzaa
393.895	1	zzgam
393.896	1	zzgamma

As can be seen, even after removing the stopWords, it is possible to verify that among the most frequent words, the majority are in Portuguese, with some of these words in their English version, such as the two most frequent words. In the last positions, however, there are words with a very low frequency. It is clear that these words have no semantic content, and one hypothesis for the existence of such words is typing errors when registering the title of the publication in a given CV. It is also clear that among the most frequent words, there are topics that are usually part of the titles of publications, since they are important for indicating methods, techniques, objects or locations.

In order to evaluate the set of words that are linked to the publications of articles in open access journals, Zipf's Law was used, as presented in Section 2.2. In Quoniam's work [10], the author describes the Zipf curve, which is divided into three distribution zones:

• Zone I - Trivial or basic information: defines the central themes of the bibliometric analysis;

• Zone II - Interesting information: located between Zones I and III and shows peripheral themes, potentially innovative information. This is where technology transfers related to new themes should be considered;

• Zone III - Noise: characterized by having concepts that have not yet emerged, where it is impossible to say whether they will emerge or whether they are just statistical noise.

In this context, the set of words identified in this thesis, after all the data processing already presented, was divided into three textual distribution zones (Figure 1).

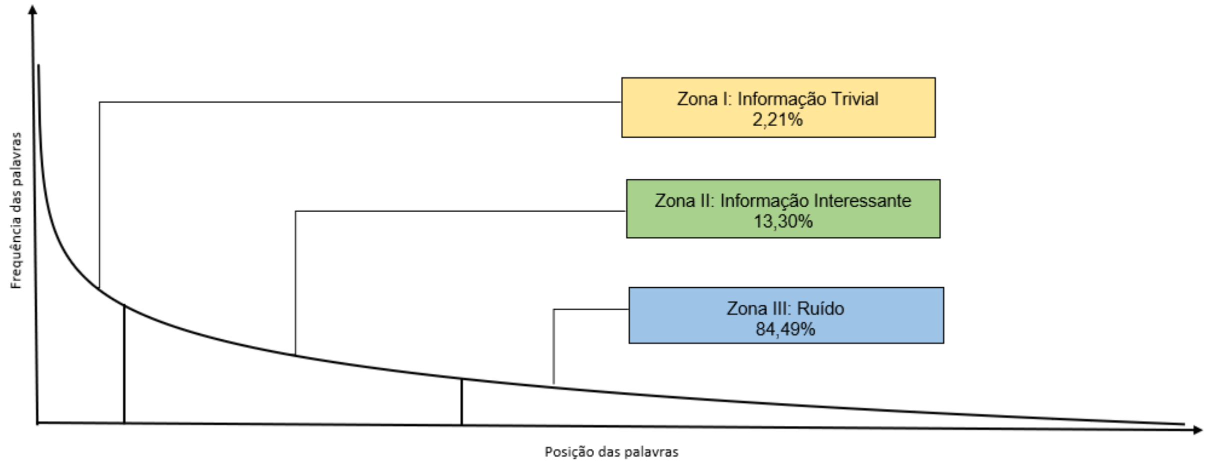

Figure 1 – Division of the Words Identified in the Three Textual Zones

The first zone identified (Zone I) has 2.21% of the words analyzed, these words being the most frequent, describing the central themes of the analyzed set. Despite having a low percentage of words, their frequency corresponds to 47.93% of the entire set, proving its representativeness. In Zone II, which has 13.3% of the words, it encompasses a set of topics that occur less frequently than those in Zone I, and because they are not words used as frequently, they are characterized as emerging themes, since they are characterized as potentially innovative information. Finally, Zone III, which has the vast majority of words (84.49%), is characterized by aggregating topics with low frequency, considered noise. Here, it is worth highlighting, as already pointed out, the problems arising from the free registration of publication data by individuals in their resumes, in which the insertion of incorrect data is real, whether due to typing errors or even coding errors when copying text from documents in different formats. A total of 149,891 words have only one occurrence.

As can be seen, there is a great disparity between the sets of words that make up each of the identified Zones. In order to better understand each of these Zones, several data analysis and visualization techniques can be applied. Figure 2 shows a word cloud from Zone 1.

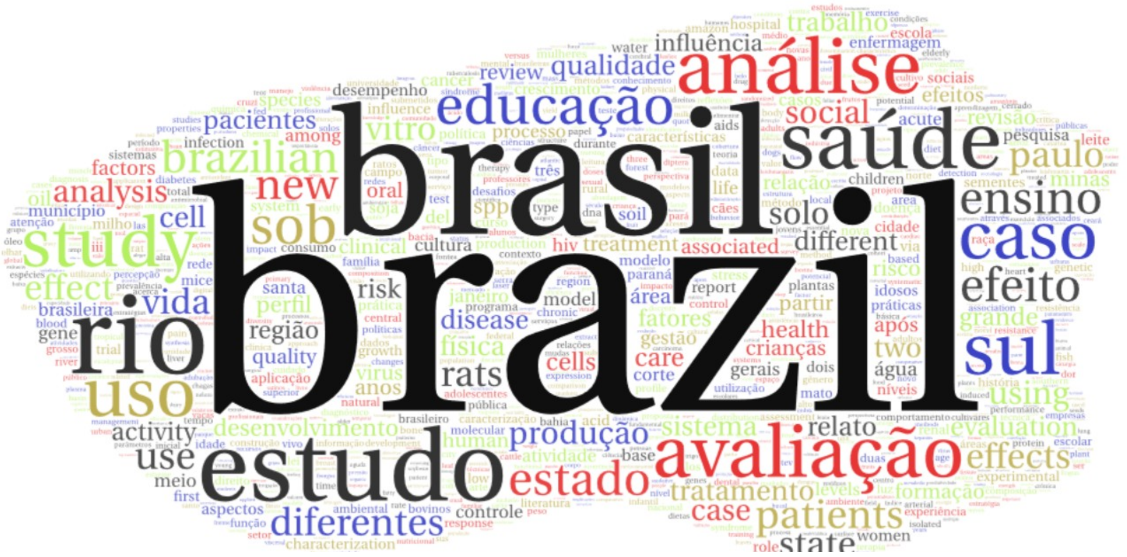

Figure 2 – Zone I: Trivial or Basic Information

It is observed that among the most frequent words in the analyzed set, the words “brazil” and “brasil” stand out significantly. As a hypothesis, it can be inferred that these words are widely used to indicate locations where the research was carried out, mainly because they consider almost all articles published by Brazilians, in English and Portuguese. In addition, the words “estudo”, “avaliação”, “análise”, “study”, and “analysis” stand out, which generally indicate methods used to conduct the research. It is also important to highlight the words “saúde” and “educação” with 58,977 and 46,863 occurrences respectively, also presenting themselves as very representative topics in the research carried out. It is worth highlighting here the occurrence of other words such as “rio”, “paulo”, “caso” and “meio” which also have significant frequency, but which may have been influenced by the method used to identify the words in the titles, considering that these are words that may also have been derived from compound words.

4 Conclusion

It is worth noting that in the study presented here, Zipf's Law was adopted in order to identify the main research topics of Brazilian researchers in publications in open access journals. To this end, all articles published in this publication were initially searched in the curriculum data repository of the Lattes Platform. The set of titles of each publication was used, and stopwords were removed from each of the titles, given their high frequency of use in article titles, which could have some impact on the analysis. In addition, it is also important to highlight that it was not possible to unify words in singular and plural, as well as the use of compound words, given that it would be necessary to adopt techniques such as Natural Language Processing, such as radicalization and n-grams, which go beyond the scope of this work.

References

[1] Mueller, S. P. (1999). O círculo vicioso que prende os periódicos nacionais (The vicious circle in which national periodicals are trapped). DataGramaZero-Revista de Ciência da Informação, (dez/99).
[2] Furnival, A. C. M., & Silva-Jerez, N. S. (2017). Percepções de pesquisadores brasileiros sobre o acesso aberto à literatura científica. Informação & Sociedade, 27(2).
[3] Hayashi, M. C. P. I. (2012). Sociologia da Ciência, Bibliometria e Cientometria: Contribuições para a Análise da Produção Científica. In: SEMINÁRIO DE EPISTEMOLOGIA E TEORIAS DA EDUCAÇÃO, 4., 2012, São Paulo. Anais... . São Paulo: Episted, 2012. p. 1 - 10.
[4] Kleinubing, L. S. (2010). Análise bibliométrica da produção científica em gestão da informação na base de dados LISA. RDBCI: Revista digital de biblioteconomia e ciência da informação, 8(2), 1-11.
[5] Cunha, M. V., Rosa, M. G., de Sousa Fadigas, I., Miranda, J. G. V., & de Barros Pereira, H. B. (2013, August). Redes de títulos de artigos científicos variáveis no tempo. In Anais do II Brazilian Workshop on Social Network Analysis and Mining (pp. 194-205). SBC.
[6] Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. Bmj, 351.
[7] Mryglod, O., Holovatch, Y., Kenna, R., & Berche, B. (2016). Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster. Scientometrics, 106(3), 1151-1166.
[8] Ronda-Pupo, G. A. (2016). Knowledge map of Latin American research on management: Trends and future advancement. Social Science Information, 55(1), 3-27.
[9] Gomes, J. O. (2018). Uma análise temporal dos principais tópicos de pesquisa da ciência brasileira a partir das palavras-chave de publicações científicas. 127 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Dezembro 2018.
[10] Quoniam, L. (1992). Bibliométrie sur des référence bibliographiques: methodologie. In: DESVALS H.; DOU, H. (Org.). La veille technologique. Paris : Dunod, 1992. p. 244-262.