A Strategy for the Identification of Articles in Open Access Journals in Scientific Data Repositories

Patrícia Mascarenhas Dias, Thiago Magela Rodrigues Dias(✉), and Gray Farias Moita

Federal Center for Technological Education of Minas Gerais, Belo Horizonte, Brazil

Abstract. This work aims to identify articles published in open access journals registered in the Lattes Platform curricula. Currently, the curricular data of the Lattes Platform has been the source of several studies that adopt bibliometric metrics to understand scientific evolution in Brazil. However, when registering a publication in a curriculum, only basic information of the journal is informed. Therefore, in order to quantify the publications that were made in open access journals, a strategy that uses DOAJ data is proposed, validating the publications and thus obtaining a process that allows identifying which publications were made in this format of communication. As a result, it was possible to quantify in an unprecedented way the set of publications by Brazilians in open access journals.

Keywords: DOAJ · Lattes platform · Data link · Open access

1 Introduction

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 bibliometric research and studies, scientific communication emerges today as a central element at various levels of discussion. Therefore, the scientific journal appears as an important mechanism for communicating research results.

[7] states that the scientific journal performs at least four essential functions: certification of science with the support of the scientific community; communication channel between scientists and wider dissemination of science; scientific file or memory and record of authorship of scientific discovery.

According to several studies, journals, mainly those available in electronic format - have been growing since the last decade. It can be said that journals, in all areas of knowledge, have the role of being a filter for the recognition of works that have been accepted. For [11], publication in a magazine recognized by the area is the most accepted way to register the originality of the work and to confirm that the works were reliable enough to overcome the skepticism of the scientific community.

In this context, in the early years of the 21st century the Open Access Movement, whose definition is "to make available to any internet user to read, download, copy,

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 125–135, 2021.
https://doi.org/10.1007/978-3-030-77417-2_9

distribute, print, search or reference the full text of articles or use them for other purposes without any barriers, as long as the work is properly recognized and cited", encouraged the appearance of journals in this format [5].

Despite the numerous benefits that open access journals provide, there is a need for a joint effort so that the main element of the whole process, scientific information, is accessible to all interested parties. To this end, some initiatives have already been undertaken, such as the creation of digital repositories to store and organize scientific literature in accordance with international interoperability standards and the search for awareness of the main actors involved in the process of production, publishing and evaluation of scientific information, to make such content available in digital environments open to the general public.

Neubert et al. [8] state that open access assumes an important role in the entire context of scientific activity, as it allows the researcher to have access to the results of other studies without the cost barriers and difficulties of access, in addition to promoting visibility and dissemination the results of the scientific activities of each researcher and each university.

Open access scientific publication is part of a broader scenario in favor of opening knowledge in general (open access, open data, open educational resources, free software, open licenses) and is essentially a movement towards the design of information and knowledge as public goods [4].

It is worth mentioning that there is generally a limited amount of resources to promote research and a large number of researchers or institutions interested in these resources. Therefore, the broader and more accurate this understanding of scientific production, the greater the possibility of determining resources correctly. However, this type of assessment is an extremely complex task, as it involves the analysis of different characteristics, both quantitative and qualitative. In addition, there is no consensus on which measures or characteristics should be considered for the assessment of scientific productivity [3].

Bearing in mind that a large part of scientific research in the country is financed with public resources, usually in public educational institutions or research centers, it is expected that the results of such studies will be disseminated without any type of barrier, mainly financial. In this context, coupled with the advantages that open access publications have, 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.

Therefore, understanding how the publications of a certain group of researchers have been carried out in open access journals, makes it possible to identify an overview of the current stage of this type of communication in Brazil. It also allows to verify if in certain areas of knowledge this type of publication tends to be more frequent.

This type of study is characterized as an important mechanism to evaluate the evolution of publications in open access journals by Brazilian researchers, allowing to verify whether the incentive policies for the publication of research in this communication format have achieved satisfactory results.

2 Related Works

In the work of [12], the authors analyzed the policies of open access to scientific information and the proposals for action, with an emphasis on government initiatives in different countries. It was identified that the movement of free access to scientific information was already a concern officially registered in several countries, although with different degrees of development. Among these differences are the policy determinations themselves, as some oblige public institutions and researchers to make their research results available in open access, while others only suggest the involvement and participation of these researchers and institutions in the movement.

[1], the main open access scientific communication channels used by researchers are identified, and the factors involved in adhering to the self-archiving of their scientific production are analyzed. The objective of the work was to identify the main channels of scientific communication in open access used by researchers from public universities in the State of Rio de Janeiro. The list of 47 CNPq Advisory Committees for Research Productivity Scholarships was used and a stratified probabilistic sampling by knowledge area was carried out, following the division by Advisory Committee (Agrarian Sciences, Biological Sciences, Exact and Earth Sciences, Science of the Health, Human Sciences, Applied Social Sciences, Engineering, and Linguistics, Letters and Arts). From the selection of researchers contemplated by the CNPq Research Productivity Scholarship program in 2010, whose list is available on the website of this federal agency, those linked to public universities in the State of Rio de Janeiro with post-graduate courses were identified. stricto sensu graduation. After identifying the e-mail addresses of those selected, correspondence was sent containing the form with closed and open questions attached to the following categories: informational behavior, open access publication and adherence to institutional repository.

In general, the results of the research point to a change in the attitude of these researchers in relation to the publication of research results in open access channels. Some areas present publications in formal channels of scientific communication, such as electronic journals, and self-archiving in institutional or thematic repositories. Others are more part of individual or group research initiatives, often anticipating institutional policies. The researchers were unanimous regarding the advantages of open access publishing, and the democratization of knowledge was pointed out by the majority as the main advantage of this adhesion. In addition to this aspect, the benefit of communication between peers – "exchanges", "partnerships" and "dialogues" - also appears in the speeches of the researchers in the knowledge production process. It is also signaled the importance of using this open communication channel at two different times: for the researcher to access the information for their research and to make their results available, allowing them greater visibility and impact.

In order to explore the national and international scenario and thus present an investigation that seeks a technological solution to effect open access to research data, [10] propose a methodology divided into five stages: a) identification of practices of open access to research data in Brazilian institutions; b) mapping your users and their needs; c) proposal for a web portal to bring together the national community; d) survey of services and technological solutions existing in the international scenario for the sharing of research data; e) proposing recommendations to support the creation of research data

repositories in national institutions and their aggregation to a research network with open access to research data. As a result, international initiatives and strategies are proposed for the creation of a research data repository and for the creation of communities of practice around the subject.

For [6], the spread of the open access movement in Latin American and Caribbean countries, driven by the growth of regional and national initiatives such as the creation of digital magazine libraries in open access and the establishment of government policies of support, has provided evidence of the significant role of open access for the participation of these countries in global scientific production. In the work, open access publications from Latin American and Caribbean countries are mapped, through a bibliometric analysis of the publications indexed by WoS and SciELO during the period from 2005 to 2017. It is found that the publications have intensified significantly in the period examined, and that although there is an increase in the number of publications in this format, in some countries its magnitude does not translate into a relative weight of open access in the total number of publications.

In the work of [9], the authors analyze documents published in open access between the years 2012 and 2016 by authors with Brazilian affiliation and identify the profile of these publications. For this, data from 930 journals and 63,847 documents were collected from WoS. It is also noteworthy that the Brazilian scientific production in open access is characterized by an endogenous profile, and that policies are still necessary to encourage the publication of articles in open access, mainly in international journals.

Considering the works that analyze publications in open access journals, it is clear that most of them analyze small sets of individuals, in addition to using international data repositories, thus neglecting publications from some areas of knowledge and, therefore, not representing significantly the Brazilian production in open access as a whole.

The vast majority of studies that evaluate the open access movement do not have publications in this type of format as their main object of study, but repositories or journals in open access. Therefore, although the works presented in this section are important to understand the existing initiatives and opinions of Brazilian researchers, as well as the main open access repositories in Brazil. A comprehensive study of Brazilian researchers who have published widely published papers in open access journals is necessary.

3 Development

For the process of data extraction for the analyzes to be carried out in the context of this work, curricular data from the CNPq Lattes Platform were used. A large part of the funding notices for research projects, carried out by various funding agencies, use data registered in the applicants' curricula as one of the forms of evaluation of the proposals. Therefore, there is a great incentive for researchers to keep their curriculum information up to date. This makes Lattes Platform curricula an excellent source of data for analysis. For this same reason, several works have used the Lattes Platform as a data source for several studies on different topics, such as networks of scientific collaborations, analyzes of productivity, academic genealogy, among others [2, 3].

Considering that the majority of related works analyzed only specific groups of individuals, and considering that the manipulation of large amounts of curricula from

the Lattes Platform is not a trivial task, since there are problems involving information retrieval and efficient algorithms for handling large volumes of data, LattesDataXplorer [2], a framework for data extraction and treatment, developed by the research group of this work was used.

As already explained, a curriculum registered in the Lattes Platform can contain various information capable of helping to understand the evolution of Brazilian science from different perspectives. However, to serve the purposes of this work, only data from publication of articles in open access journals were considered. In view of this, an extension of LattesDataXplorer was proposed with the inclusion of non-existent a priori components, which would evaluate for each article published in a journal (6,985,179), of each of the individuals (5,901,161) (data collections in October 2018), if the journal in which that article had been published was open access (Fig. 1). Therefore, with the proposal for this extension, only authors and publications in open access journals could be analyzed.

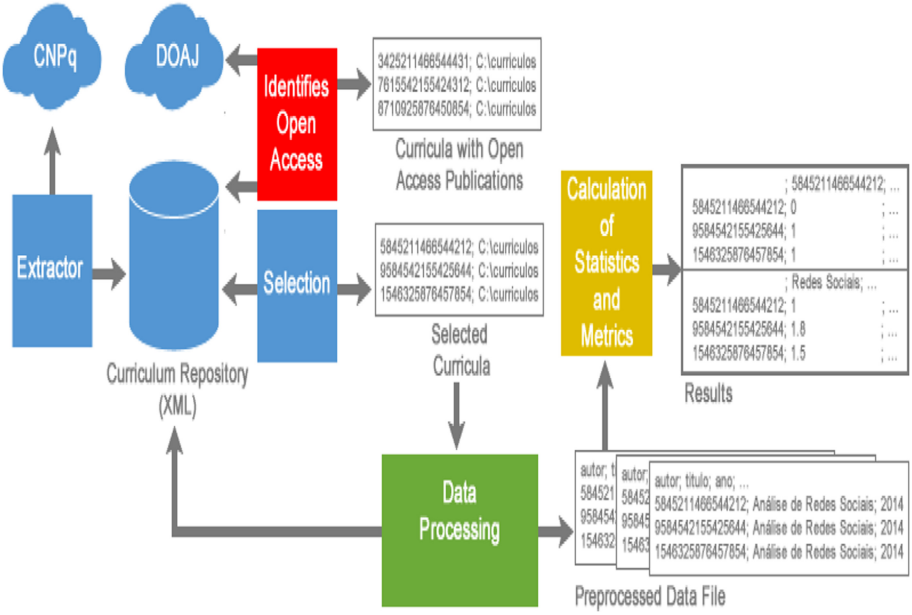

Fig. 1. LattesDataXplorer extended. Source: Authors.

Initially, using LattesDataXplorer, all resumes registered on the Lattes Platform in October 2018 were collected and stored in the local repository. Then, the component developed and called "Identifies open access" was used to retrieve all open access journals registered on the Directory of Open Access Journals (DOAJ) portal, an online directory that indexes and provides access to open access journals. In February 2019, the DOAJ indexed 12,324 journals and 3,513,782 articles. DOAJ has been a source of data and reference on open access journals for several studies.

Collected the data of the journals on the DOAJ portal in October 2018, the same period of collection of the curricula for the analyzes presented in the present work, 12,171 open access periodical titles were retrieved, containing data such as title, ISSN and eISSN, among other information.

In order to optimize the computational processing of curricula as much as possible, whenever a publication whose ISSN or eISSN of the journal was contained in the list of open access journals extracted from DOAJ, immediately the identifier of the curriculum under analysis was inserted in the list of curricula in access. open, and the next curriculum of the set under analysis was evaluated.

After analyzing all the resumes that make up the local repository, a list containing all resumes with open access publications is generated, and it becomes the basis for the "Data processing" component, which now incorporates the methods proposed in this work (Fig. 2).

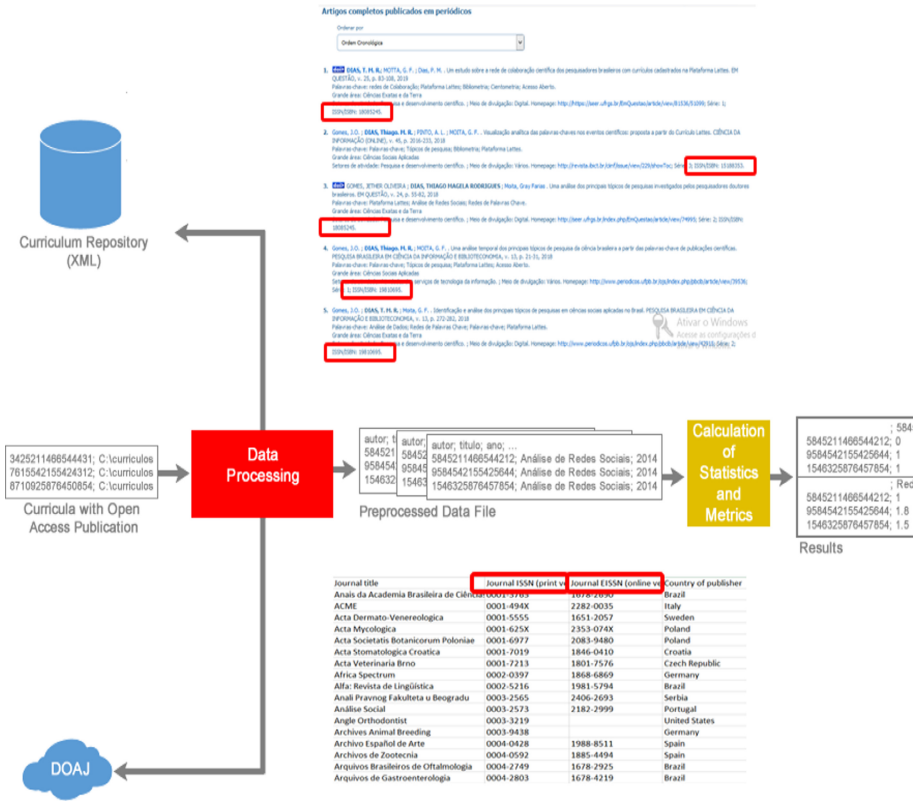

Fig. 2. Method for identifying publications in open access journals. Source: Authors.

With the list of curricula that have articles in open access, the identification of publications in this format is performed with the processing of the curricula, using the "Data Processing" module of LattesDataXplorer, in order to generate the pre-processed

data files that summarize information of interest and that will serve as the basis for calculating the metrics.

In addition to general data on researchers with open access publications that will compose some of the archives, such as data on academic training, areas of expertise, guidelines and professional practice, each of the articles recorded in the section "Complete articles published in journals" was analyzed. Of each curriculum contained in the "List of curricula with publications in open access". For each article in each curriculum, it was verified and analyzed whether the ISSN or eISSN of the publication was present in the list of journals recovered from DOAJ. Thus, it was possible to identify the entire number of articles in open access journals (Fig. 3).

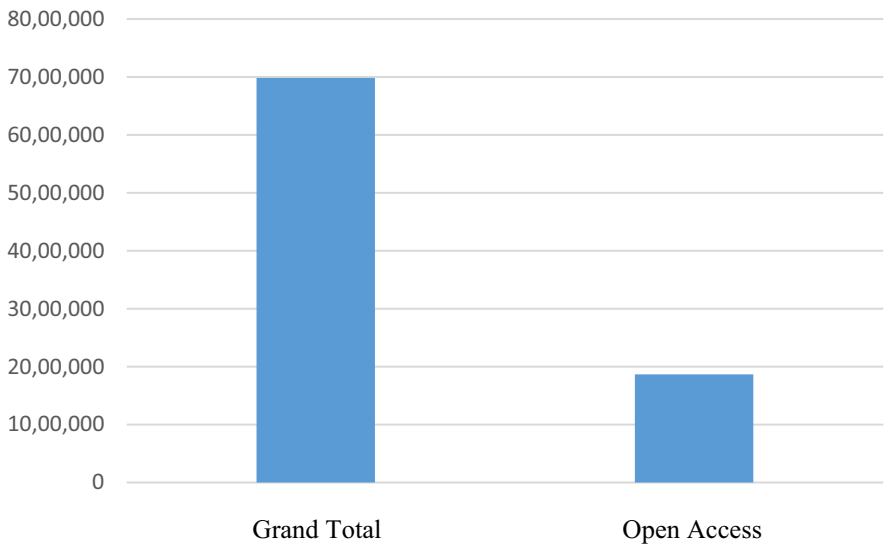

Fig. 3. Number of publications in journals registered in the curricula. Source: Authors.

As can be seen, of the total set of articles published in journals, considering the entire history of publications recorded in all resumes registered in the Lattes Platform (6,985,179 publications), a percentage of 26.76% (1,869,585) was published in open access journals, taking into account the list of journals recovered from DOAJ. This percentage of publications in open access is relevant, above all, for considering the entire publication history of each researcher. It is noticed that publications in open access journals have been receiving attention and adherence by researchers year after year, presenting themselves as a trend in dissemination and scientific communication, especially in recent years. A temporal evaluation was carried out in order to assess the growth of publications year by year.

4 Results

Using the extension proposed in this work for LattesDataXplorer, all authors who published at least one article in an open access journal (370,431) were identified. These

authors, despite being a small number of individuals in relation to the whole set registered in the Lattes Platform (6.27%), have a great representativeness when considering the total number of articles published in journals (approximately 76%) (see Fig. 4).

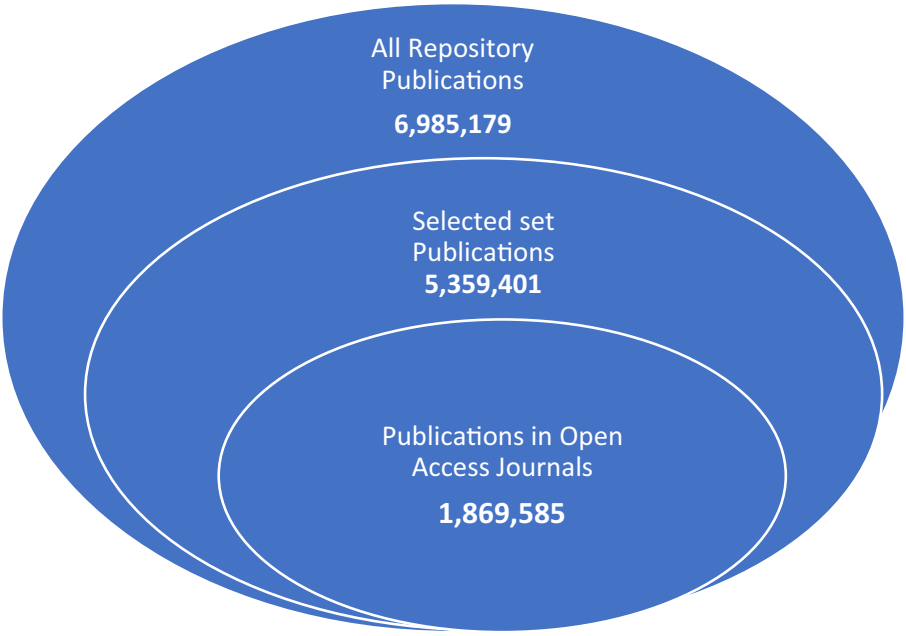

Fig. 4. Number of publications in registered journals. Source: Authors.

Therefore, it is possible to note the representativeness of the set to be analyzed in this work. Bearing in mind that it includes a considerable portion of the authors who have published articles in journals in Brazil, the results presented may provide an unprecedented view on the evolution of articles in open access, as well as serve as a basis for several other works.

It is possible to inform in the curricula the areas, subareas and specialties in which a given individual operates. When analyzing the areas of activity of the group of individuals, it is possible to notice great diversity in the distribution of curricula in each major area, as well as an irregular distribution in the number of areas that each major area has. Therefore, an analysis based on the areas of activity is important (Fig. 5).

As can be seen, there is no uniform distribution of the number of areas in each large area. The large area of Linguistics, Letters and Arts has only three areas, while the large areas of Biological Sciences and Engineering have 14 areas each. When registering the large areas of activity, the individual may not inform the "area" field. In these cases, individuals were also categorized as "Not Informed". Due to the small number of individuals (0.81%) who reported "Others" as a large area, the analysis of their areas was not considered.

In order to verify the most representative areas of knowledge of the analyzed group, the area of Medicine (33,966) stands out, composing the large area of Health Sciences,

Fig. 5. Distribution of authors by their areas of expertise. Source: Authors.

the most representative, as already presented. The representativeness of the field of Medicine is so considerable that it alone has practically the same number of individuals as the total sum of the major areas of Linguistics, Letters and Arts and Engineering combined. Following, the areas of Education (17,066), Agronomy (15,705), Nursing (12,734), Administration (11,717), Chemistry (10,490) and Psychology (10,511) stand out. These seven areas alone are responsible for housing approximately 30% of the total set of individuals analyzed.

5 Final Considerations

In order to draw a picture of the publication of articles in open access journals by Brazilian researchers, it was necessary to develop components that, incorporated into LattesDataXplorer, could enable the analyzes carried out in this project. Thus, the entire curriculum data repository of the Lattes Platform was analyzed, enabling an unprecedented study on the Brazilian production of articles in open access journals using data from DOAJ as well.

The set of articles published in open access journals has as authors a total of 370,431 individuals, which represents approximately 6% of the total set of individuals with curricula registered in the Lattes Platform. It should be noted that this percentage of authors is much lower than the number of articles in open access journals, which represent approximately 27% of the total number of articles published in journals of all individuals. This percentage is very close to that presented by [4], who point out that only around 30% of the total scientific articles published in the world annually are available through open access channels. Therefore, it is identified here that the percentage of publications in open access journals in Brazil is slightly lower than the world average of publications in this format. This study will provide several new researches that aim to broadly analyze the production of articles in open access journals in Brazil.

References

1. Chalhub, T., Pinheiro, L.V.R.: Acesso aberto à informação científica no Brasil: Um estudo das universidades públicas do estado do Rio de Janeiro. Rio de Janeiro (Relatório Final de Atividades) (2011)
2. Dias, T.M.R.: Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes. 2016. 181 f. Tese (Doutorado) - Curso de Programa de Pós-graduação, Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte (2016)
3. Digiampietri, L.A.: Análise da Rede Social Acadêmica Brasileira. 2015. 160 f. Tese (LIVRE DOCÊNCIA) - Informação e Tecnologia, Escola de Artes Ciências e Humanidades da Universidade de São Paulo, São Paulo (2015)
4. Furnival, A.C.M., Silva-Jerez, N.S.: Percepções de pesquisadores brasileiros sobre o acesso aberto à literatura científica. Percepções de Pesquisadores Brasileiros Sobre O Acesso Aberto à Literatura Científica, João Pessoa, vol. 27, no. 2, pp. 153–166 (2017
5. Leta, J., Costa, E.H.S., Mena-Chalco, J.P.: Artigos em Periódicos de Acesso Aberto: um Estudo com Pesquisadores Bolsistas de Produtividade do CNPq. Revista Eletrônica de Comunicação, Informação e Inovação em Saúde [s.l.], vol. 11, pp. 1–6 (2017)
6. Minniti, S., Santoro, V., Belli, S.: Mapping the development of Open Access in Latin America and Caribbean countries. An analysis of Web of Science Core Collection and SciELO Citation Index (2005–2017). Scientometrics **117**(3), 1905–1930 (2018). https://doi.org/10.1007/s11192-018-2950-0
7. Mueller, S.P.M.: O círculo vícios o que prende os periódicos nacionais. Datagramazero, Brasília, vol. 0, no. 4, pp. 1–8 (1999)
8. Neubert, P.S., Rodrigues, R.S., Goulart, L.H.: Periódicos da Ciência da Informação em acesso aberto: uma análise dos títulos listados no DOAJ e indexados na Scopus | Open access journals in information Science. Liinc em Revista, [s.l.], vol. 8, no. 2, pp. 389–401. Liinc em Revista (2012). https://doi.org/10.18617/liinc.v8i2.497

9. Pavan, C., Barbosa, M.: Article processing charge (APC) for publishing open access articles: The Brazilian scenario. Scientometrics **117**(2), 805–823 (2018). https://doi.org/10.1007/s11192-018-2896-2
10. Pavão, C.G., Rocha, R.P., Gabriel Junior, R.F.: Proposta de criação de uma rede de dados abertos da pesquisa brasileira. Rdbci: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, vol. 16, no. 2, pp. 329–343 (2018)
11. Rodrigues, R.S., Oliveira, A.B.: Periódicos Científicos na America Latina: títulos em Acesso Aberto indexados no ISI e SCOPUS. Perspectivas em Ciência da Informação, Belo Horizonte **17**(4), 76–99 (2012)
12. Silva, T.E., Alcará, A.R.: Políticas de acesso aberto à informação científica: iniciativas governamentais. In: IX Encontro nacional de pesquisa em ciência da informação, 9, São Paulo. Anais. São Paulo, pp. 1–14 (2008)