Collection and Processing of Data from Articles in Scientific Events

Fernanda Coimbra[1], Thiago M. R. Dias[1], Patrícia M. Dias[2], Guilherme M. Dias[3]

[1]*Federal Center for Technological Education of Minas Gerais*
Rua Álvares de Azevedo 400 - Bela Vista – Zip-Code: 35503-822 - Divinópolis - MG - Brazil
fernandacoimbrastilo@gmail.com, thiagomagela@cefetmg.br
[2]*State University of Minas Gerais*
Av. Paraná, 3001 - Jardim Belvedere – Zip-Code: 35501-170 - Divinópolis – MG - Brazil.
patriciamdias@gmail.com
[3]*Federal Institute of Education, Science and Technology of Minas Gerais*
Rua Padre Alberico, nº 440, bairro São Luiz – Zip-Code: 35577-020 - Formiga – MG - Brazil.
guilhermedias2501@gmail.com

Abstract. Several studies focus on exploring the behavior of scientific evolution. For this reason, one of the main means of scientific communication today is events, in which diverse and valuable works are generated, also enabling fast communication and possibilities for argumentation. However, most works that evaluate scientific production in scientific events generally have specific repositories as a data source, often restricted to some areas of knowledge. Although such works present interesting results, no studies were found that encompass events in general, with a large number of publications or that address different areas. In this context, this work aims to analyze the scientific production published in the annals of events of the group of physicians who have CVs registered on the Lattes Platform. The data used were extracted from the Lattes Platform in January 2021, so that all the selection and processing of the data of interest could be carried out. Therefore, this work has the general objective of analyzing the scientific production published in events, understanding how Brazilian scientific production occurs in this means of dissemination, using bibliometric metrics in the data extracted from the curricula registered in the Lattes Platform.

Keywords: Scientific production, Events, Lattes Platform.

1 Introduction

The Internet has become a large repository of scientific knowledge, providing users with access to this valuable asset in a simple and intuitive way, allowing them to make available and access scientific works, which include their analyses and results, as well as enabling these users to disclose their personal, professional and academic information. Currently, there are several sites for recording this information and several institutional repositories that also enable scientific dissemination by groups of individuals.

There are numerous studies available on the Internet that use data from scientific productions, and works that analyze scientific publications have gained prominence. According to Domingues [1], scientific productions are an integral part of the individual's knowledge production process, in which this knowledge acquired can be made available through articles in conference proceedings and in periodicals, books, book chapters, abstracts, theses, dissertations and monographs, among other means of disseminating and communicating science.

The volume of data currently available has its own characteristics, unique patterns, quantity and diversity of

data, thus making it a complex task to conduct studies that aim to explore this data. In this context, bibliometrics emerged with the aim of quantifying written communication processes, through the use of methods that generate statistical analyses on the production and dissemination of knowledge applied to scientific data sources [2]. Currently, there are several quantitative techniques to assess scientific productivity, such as bibliometrics, scientometrics, informetrics and webometrics. All techniques allow for different analyses. In his study, Dias [3] explains that bibliometric analyses can be used to measure the scientific production of particular individuals, research groups, institutions, geographic regions, organizations or events.

Over the last few years, it has been possible to verify the evolution of scientific dissemination in the various areas of knowledge and in this context, it has been possible to verify how this increase has been driven by the increase in scientific events in the world. Classified as one of the main means of scientific communication, events, considered an effective means of oral communication of knowledge, have been gaining notoriety [4].

For the authors Campello, Cédon and Kremer [5], scientific events perform functions such as: improvement of works, reflection of the state of the art and communication. In scientific events, documents are generated with the works presented, popularly known as annals. The publications generated in the events are considered by some studies as the most current academic productions [6]. The behavior of scientific publications in a given field of research allows new perspectives for understanding it, enabling a new approach in the evaluation of science [7]. In view of this, analyzing how publications in event annals have been carried out is presented as an important mechanism for understanding the evolution of scientific events in a general context or in certain areas of knowledge. However, in general, information related to scientific production in event proceedings is present in numerous data repositories, thus making data retrieval and analysis difficult, especially on a large scale. There are several studies that analyze events in specific repositories or by specific area.

As in the study by Fathala et al. [8], which aims to obtain a better understanding of the characteristics of academic events in four fields of science; they analyze the metadata of academic events from four major fields of science; renowned academic events belonging to five subfields; through the analyses, the authors find expressive results that allow observing the general evolution and success factors of academic events, thus allowing event organizers to judge the progress of their event over time and compare it with other events in the same field; analyses are also presented that enable decision-making for researchers to choose the appropriate locations to present their work.

In order to obtain a large set of data on scientific publications in conference proceedings, the Lattes Platform of CNPq (National Council for Scientific and Technological Development) emerges as an excellent alternative for data collection. One of the elements of the Lattes Platform is the CV, in which it is possible to include and retrieve information entered in the CVs by the individuals themselves; it is characterized as an open access platform. In this context, in Brazil, the Lattes Platform has become a standard for registering CV data for the scientific community. According to Lane [9], the Lattes Platform is a powerful Brazilian scientific data repository, which has high-quality data and allows access to the data of registered individuals. It is the individual's own responsibility to enter and update the data. Because it is a valuable repository, it is possible to find in the literature several authors who use the Lattes Platform as a source of bibliometric studies. Dias [3] presents the first work in the literature, since in addition to a broad study on Brazilian scientific production using all CVs registered on the Lattes Platform as a data source, the author develops a framework, implementing bibliometric techniques and metrics based on social network analysis, being responsible for extracting the entire set of CV data.

There are numerous studies that use the Lattes Platform as a data source resource, as it is a rich repository, including articles published in events. However, there is a limited number of studies that use curricular information from the Lattes Platform as the main data source in the context of studies published in event proceedings. Retrieving studies published in events is a complex task, due to the range of events, formats, among other points. In view of this, this study presents itself as an analysis aimed at understanding the behavior of studies published in event proceedings registered in the curricula registered in the Lattes Platform in a global manner, considering all registered curricula, through bibliometric analyses, general characterization of events, temporal analyses, and quantitative analyses, making it possible to verify representativeness by areas of knowledge.

There are several points that motivate us to understand scientific production through conference proceedings. One of them is the possibility of an overview of how different areas of knowledge have explored this means of dissemination to present the results of their research; in which time periods there was a greater

number of publications and which areas have the highest volume of publications. In view of this, this work has the general objective of understanding features of Brazilian scientific production in conference proceedings, using articles registered in the Lattes Platform CVs as a data source.

2 METHODOLOGY

The Lattes Platform integrates databases into a single system for CVs, research groups and institutions. However, it is the individual's responsibility to insert all their CV information into the Lattes Platform, and after inclusion, all this data is available in open access on the internet. It is a rich repository, including records of professional and academic careers and scientific production, which allows for several different analyses, thus justifying the choice of this repository as a data source for this work. However, it is not possible to retrieve all CVs at once.

The methodology used in this study was based on bibliometric analysis concomitantly with quantitative methodology. LattesDataXplorer [3] was used; this framework allows for the process of extracting and selecting CV data from the Lattes Platform, which involves a set of techniques and methods that enable the collection, selection, processing and analysis of data. For this work, only the collection and selection modules of LattesDataXplorer were used to extract and select the curriculum data from the Lattes Platform. Thus, the Collection module was carried out in stages, namely: collection of URLs (Uniform Resource Locator), collection of identifiers and extraction of the curriculums.

The LattesDataXplorer extractor was used in January 2021 to collect all the curriculums, approximately 7 million records. The curriculums are in XML (eXtensible Markup Language) format, which allows for delimitations and is suitable for automatic processing, enabling better data manipulation.

To obtain an accurate analysis of the articles published in event proceedings, it was preferred to establish a set of individuals by level of academic training/degree who have completed a doctorate. The choice is made based on what Dias [3] mentions in his work: PhDs are responsible for 74.51% of articles published in journals and 64.67% of articles published in conference proceedings, in addition to generally having recently updated their CVs and notably being responsible for the highest level of education. In order to assist in data analysis, after using the framework, it was necessary to create methods for selection, processing and visualization (Figure 1).

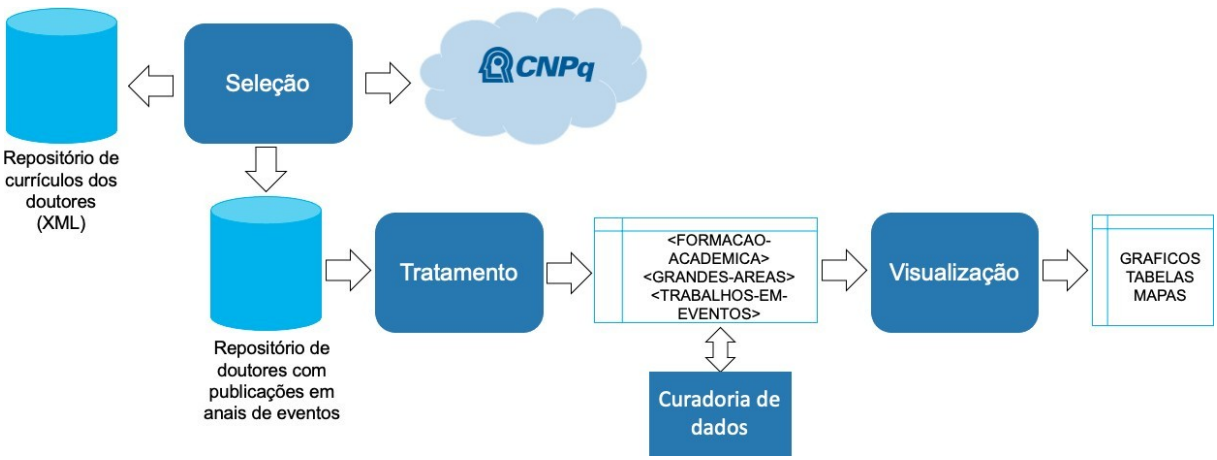

Figure 1: Data selection, processing and visualization process.

The processing stage was based on analyzing the set of XML files. For each resume, a file in XML format was extracted. The XML resume extracted from the Lattes Platform presents its structure as the root element, called “Resume”; and has five child elements that have their own elements and attributes. Each resume is unique and has its own information; this data aggregates information about major areas, academic background,

guidance, productions, among others. In this stage of the analysis, the resume information is divided, then the information of interest is accessed and some information that is irrelevant for this work is discarded. After processing the XML, the visualization stage was carried out in which the data is characterized, allowing the analysis of the data entered in the Event Papers Section, a section that contains papers published in event proceedings, if the individual has informed this in his/her resume. Through data characterization, it was possible to obtain general indicators such as: total number of articles published in event proceedings, year of publication of articles, total number of individuals who have articles published in event proceedings, publications by major area, individuals without publications, publications with persistent identifiers; such indicators are presented in the Results section.

3 Results

The initial characterization resulted in a set of 360,888 CVs of individuals with completed doctorates, this amount represents approximately 5% of CVs registered on the Lattes Platform (data from January 2021). Among these CVs, it was found that 57,403 CVs did not have any articles listed in the section of works published in conference proceedings, corresponding to 16% of the set of PhDs. After analyzing the data, it was possible to understand how Brazilian scientific production occurs in conference proceedings using the CVs registered on the Lattes Platform as a data source, which can be analyzed by major areas of activity and perform a temporal analysis of the publications. Since the focus of this work is to characterize the articles published in conference proceedings, for the next analyses only the data set corresponding to the 11,416,655 articles published in conference proceedings identified in this study will be used. After analyzing the data, it was possible to understand how Brazilian scientific production occurs in conference proceedings, using the CVs registered in the Lattes Platform as a data source, which can be analyzed by major areas of activity and perform a temporal analysis of the publications. Since the focus of this work is to characterize the articles published in conference proceedings, for the next analyses only the data set corresponding to the 11,416,655 articles published in conference proceedings identified in this study will be used.

The temporal analysis of articles published in event proceedings was carried out with the aim of verifying the number of articles published per year, a 30-year cut-off line was drawn for better graphic representation (Figure 2).

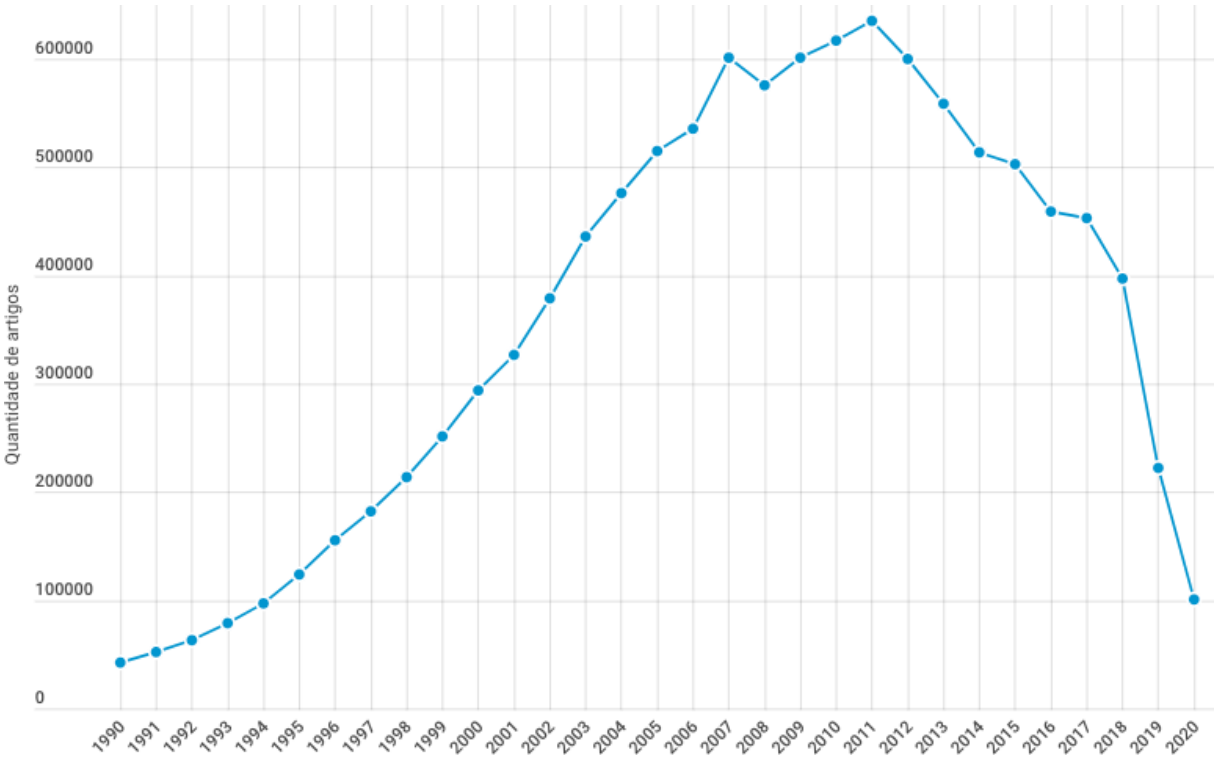

Figure 2 - Temporal analysis of articles in event proceedings by year.

It is possible to observe a constant increase in the number of publications, especially from the end of the 1990s, with the peak of publications in 2011. After that, a significant drop in publications in events was observed from 2011 onwards, and in 2018 there was an abrupt drop. However, one hypothesis for the significant drop in the number of publications in recent years may be related to the lack of updating of some CVs, which even if the author has published a work, may not have yet registered the article in his/her CV on the Lattes Platform.

The Lattes Platform CVs allow the insertion of information on the individuals' major areas of activity, following the CNPq classification of areas. Thus, the individual inserts his/her classification into nine major areas, which are: Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences, Human Sciences, Applied Social Sciences, Engineering, Linguistics, Literature and Arts, and Other. If the individual does not provide this information, it will be blank and, in the context of this study, it was characterized as “Not provided”.

In order to understand in which major areas the individuals considered in this study are distributed (Figure 3), it is worth noting that, since it is possible to include more than one major area of activity in a CV, when this phenomenon occurred, the first record was considered to be the main major area.

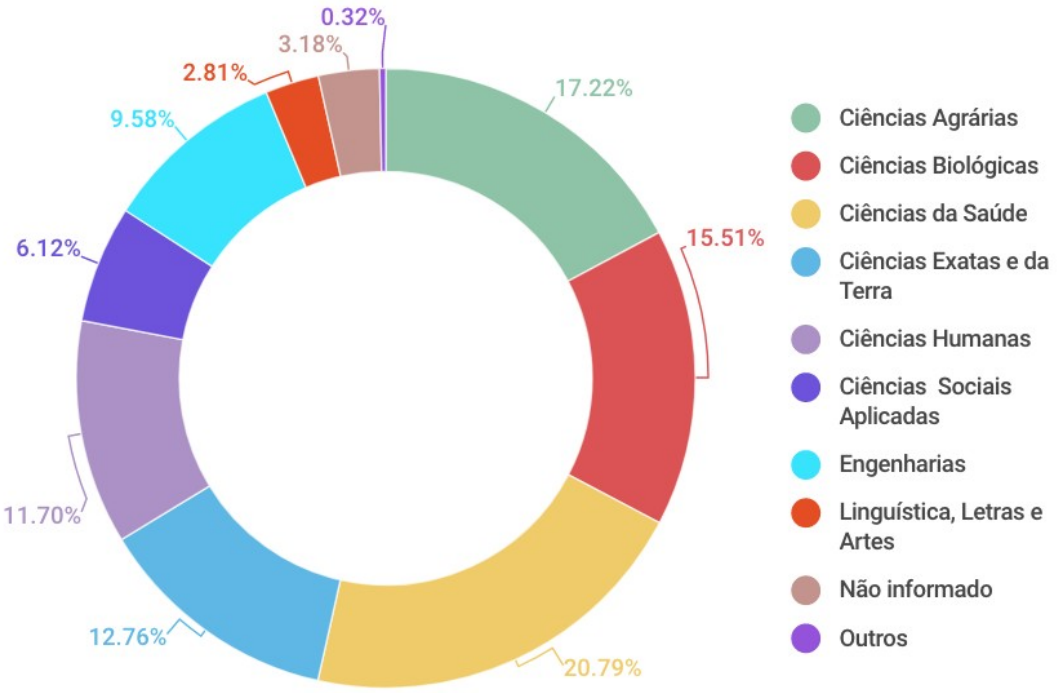

Figure 3 - General quantity of articles by major areas.

It can be seen that the major area of Health Sciences has the highest rate of publication in conference proceedings (20.79%). The major area of Agricultural Sciences (17.22%) has the second highest percentage, and a similar number to the major area of Biological Sciences (15.51%). It was also possible to observe that the major area with the highest reference value without publications in conference proceedings corresponds to Human Sciences (16.67%), followed by the major area of Applied Social Sciences (14.94%) and Exact and Earth Sciences (14.22), respectively, with similar numbers.

All data on papers published in events are entered manually by the individuals themselves. However, there is an alternative way to enter such data partially automatically, using the persistent DOI identifier. When the identifier is entered, the query is performed and the data is automatically indexed in specific fields in the CVs.

In 2007, CVs on the Lattes Platform began to accept the DOI, allowing it to be entered manually, allowing the Platform to perform a query and fill in the publication data automatically.

When analyzing the data on publications registered in the CVs of PhDs that have the persistent identifier (DOI), it is possible to verify that only 30,936 of the articles in event proceedings registered in the CVs of the selected set have a persistent identifier, thus representing approximately 3% of the articles. This shows that it is still a little-used identifier, considering that its use could aggregate a greater amount of data regarding the articles that are being registered. Thus, when recovering information that has a DOI registered in their CV productions, it was possible to distribute these across large areas (Figure 4).

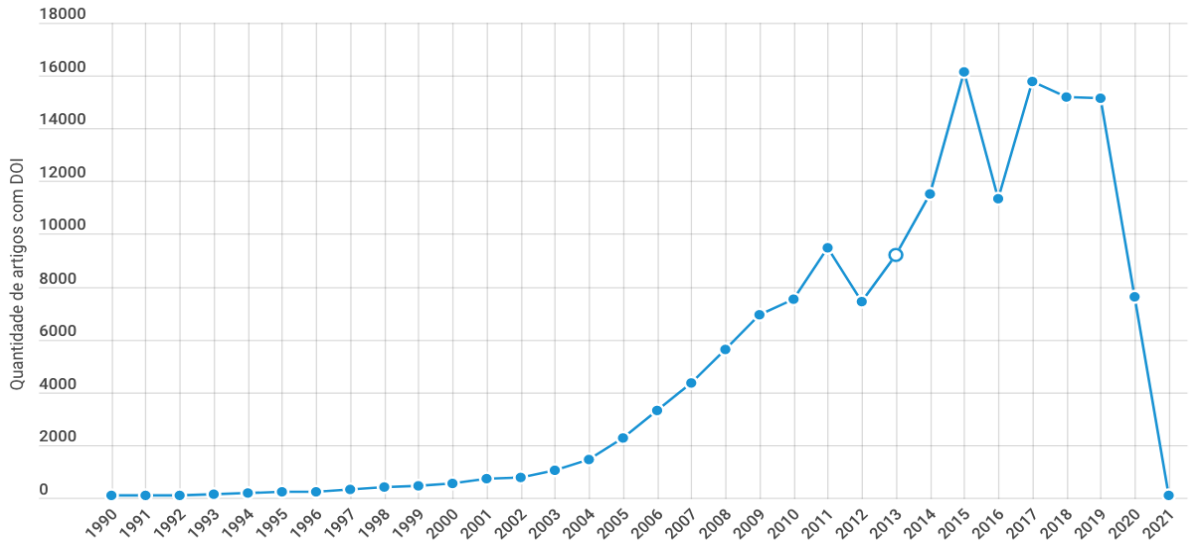

Figure 4 - Temporal analysis of articles with persistent identifier.

It is observed that there is a significant temporal variation in the use of the identifier. One of the hypotheses for this variation is the volume of events in that year, in which the number of publications may have influenced the quantities. Or, even the way in which these proceedings were published, considering that some events, under the responsibility of some organizers, value the incorporation of the DOI in the proceedings of their publications.

4 Conclusion

This study revealed that data extracted from the Lattes Platform is an excellent source for understanding how Brazilian scientific production occurs in conference proceedings. Studies that focus on analyzing articles published in conference proceedings in general are relevant and are capable of providing various analyses, in order to understand the behavior of individuals participating in events, as well as the publications that are being made. Through the analyses performed, it is possible to observe some general characteristics of the data, such as: participation in events by large area, seasonality of publications and the scarce use of persistent identifiers, countries with the highest number of participations in events.

References

[1] Domingues, I (1995), “O sistema de comunicação da ciência e o taylorismo acadêmico: ques-tionamentos e alternativas,” Estudos avançados, vol 28.
[2] Araújo, C. A. (2006), “Bibliometria: evolução histórica e questões atuais”, vol 12, 11-32.
[3] Dias, T. M. R. (2016). Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte.
[4] Mello, L. L. C. C. (1996). “Os anais de encontros científicos como fonte de informação: relato de pesquisa”, vol 20, 53-68.
[5] Campello, B. S.; Cedón B. V.; Kremer J. M. (2007), “Fontes de informação para pesquisadores e profissionais”, UFMG, Belo Horizonte.
[6] Carmona, I. V.e Pereira, M. V. (1994), “Ciência, tecnologia e sociedade e educação ambiental: uma revisão bibliográfica em anais de eventos científicos da área de ensino de ciências”, Revista Ciências & Ideias, vol 8, 94–114.
[7] Araújo, R. F., Alvarenga L. (2011), “A bibliometria na pesquisa científica da pós-graduação brasileira de 1987 a 2007”, vol 16, 51-70.
[8] Fathala S. et al (2020), “Scholarly event characteristics in four felds of science: a metrics-based analysis”, vol 123, 677-705.
[9] Lane, J. (2010), “Let’s make science metrics more scientific”, vol 464, 488-489.