Certification and characterization of technical production registered on the Lattes Platform

Raulivan Rodrigo da Silva[1], Thiago Magela Rodrigues Dias[1], Washington Luís Ribeiro de Carvalho Segundo[2]

[1]*Teacher of basic, technical and technological education, Federal Center for Technological Education of Minas Gerais (CEFET-MG)*
Divinópolis, Minas Gerais, Brazil.
raulivan@cefetmg.br, thiagomagela@cefetmg.br
[2]*Technical Coordinator of the Treatment, Analysis and Dissemination of Scientific Information Area, Brazilian Institute of Information in Science and Technology (IBICT)*
Brasília, Distrito Federal, Brazil.
washingtonsegundo@ibict.br

Abstract. The Lattes Platform is a valuable source of information, however, due to the large volume of data and filling in the information being the responsibility of the individual, it may eventually cause inconsistencies in the entered data, which emphasizes the need for validation mechanisms. Methodology: The methodological process was divided into two parts, in which the first is characterized by the collection and construction of the local database with data from patents filed in Brazil and curricula from the Lattes Platform, and the second part is the description of the patent certification process informed in the CVs of the Lattes Platform. Results: A local database was built, consisting of 903,979 patent records from Espacenet and 76,619 patent records distributed in 28,581 CVs collected on the Lattes Platform, organized in a relational database. It was possible to certify approximately 58% of the patents reported on the Lattes Platform. Conclusion: About 1% of the CVs on the Lattes Platform contain information on patents, a base composed of more than 7 million CVs (2022). Of this amount, not all could be validated and certified on Espacenet due to inconsistency in the recorded data, highlighting the need for validation and certification mechanisms for patent data.

Keywords: patent; Lattes Platform; certification; indicators.

1   Introduction

In the academic environment, it is a common practice to publish scientific articles, and it may even be considered the main way in which a large part of scientific knowledge is disseminated to the community. The publication of articles is so important in the academic scenario that it is also used as a metric to evaluate the productivity of researchers and used as an evaluation criterion in several competitions and selection processes (FERREIRA et al. [1]).

In this context, as well as the publication of articles, patents are the main way of sharing technological knowledge, in addition to the possibility of generating financial benefits for inventors, institutions and even for the country, directly contributing to technological development. In the same way as the publication of articles, the deposit of patents also belongs to the academic environment, since a large part of the patented inventions were generated in universities or with the participation of universities, whether public or private (RIBEIRO; FREY; AZEVEDO [2]).

In this context, as a mechanism for recording technical and scientific production, there are the curricula of the Lattes Platform. Which is maintained by the National Council for Scientific and Technological Development (CNPq) with the objective of integrating academic curricula bases from public and private institutions into a single platform (LOPES [3]).

1.1 Problem definition

Based on what was presented, the Lattes Platform can be considered a valuable source of information for the most diverse types of analysis, however, due to the large volume of data and filling in the information being the responsibility of the individual, it may eventually cause inconsistencies in the entered data (DIGIAMPIETRI et al. [4]). This emphasizes the need for validation mechanisms for the information contained in the Lattes Platform curricula, in order to contribute to the consistency of this important database. As well as the certification process of Theses and Dissertations informed in the curricula of the Lattes Platform carried out through Oasisbr - Brazilian Portal of Publications and Scientific Data in Open Access (BRASIL [5]).

1.2 General objectives

Therefore, this work has as main objective to propose a strategy for certifying patents informed in the curricula of the Lattes Platform in an international patent repository.

2   Methodological procedures

The methodological process was divided into two parts, the first consisting of the construction of a local data repository, and the second, the definition of the patent certification process.

2.1 Writing style

Please limit your paper by writing concisely, rather than by reducing figures or tables to a size at which symbols or labels would become difficult to read.

For the construction of the local data repository, a priori it was defined the collection of data from patents deposited in Brazil made available in the Espacenet database. Espacenet is a broad and consistent search service with worldwide coverage that offers free access to patent information from the years 1782 to the present day, containing data from more than 120 million patent documents from several countries (ESPACENET [6]).

Espacenet was chosen due to its ability to provide a web service that offers access to the EPO (European Patent Office) database through an API (Application Programming Interface), called OPS (Open Patent Services). This makes it possible to develop computational algorithms that enable more efficient consultation and data collection.

Therefore, based on the OPS documentation[1], computational routines were implemented to collect the data of interest via API, considering the patent filing period from January 1900 to December 2022.

As a result, the search API returns a list containing the filing number of the patents that met the filters. With this list in hand, for each patent on the list, a request was made to the "publication" service, informing the patent number as a parameter, for example "BR 9104913A", to obtain more information about the patent, such as title, abstract, classification, inventors, depositors, among others. The collection was carried out between the months of January and April of the year 2023.

Continuing the data acquisition process, the next step was to collect CVs registered on the Lattes Platform that have patent information. The process of collecting and selecting curricular data from the Lattes Platform was carried out using the LattesDataXplorer framework (DIAS [7]). The collection of CVs was carried out in January 2023.

With the data from Espacenet and the Lattes Platform, the local data repository was structured. To facilitate the organization and retrieval of information, the use of a relational database was adopted. The defined database was SQLite, as it is a database that does not need a server, in addition to not requiring any type of configuration.

After the construction of the local data repository, the certification of the patents informed in the curricula of the Lattes Platform was carried out. The certification process consists of identifying the patent informed by the researcher in his/her CV in the set of patents retrieved on Espacenet, using the deposit number and title of the patent as identification criteria. Based on the above, the following strategy was developed:

---

[1] OPS. Open Patent Services RESTful Web Services. 1.3.18. Available in: https://www.epo.org/searching-for-patents/data/web-services/ops.html.

1. Retrieve the patent filing number from the resume.
    a. Example: BR 2020 12345-9B2
2. Remove special characters such as . - # , / \ " ' ( ) e espaços.
    a. Example: BR 2020 12345-9 (B2) → BR2020123459B2
3. Remove the following terms: SIGILO, PROTOCOLO, REGISTRO, N°, INPI, EM ANDAMENTO, DEPÓSITO.
    a. Example: N° BR2020123459B2 → BR2020123459B2
4. Check if the number of remaining characters is greater than 4, if it is less than 4 characters, the process ends, considering the patent as not certified.
5. Search the patent in the database, if you find it go to step 10, otherwise go to step 6.
6. If the number starts with BR, search again disregarding the last 3 digits, if found go to step 10, otherwise go to step 7.
    a. Example: BR2020123459B2 → BR202012345
7. If the number starts with the prefixes CI, DI, MU or PI, search including BR at the beginning and disregard the last 3 digits, if found go to step 10, otherwise go to step 8.
    a. Example: PI2020123459B2 → BRPI202012345
8. If the number starts with the prefixes CI, DI, MU or PI, search replacing the prefix with BR and disregard the last 3 digits, if found go to step 10, otherwise go to step 9.
    a. Example: PI2020123459B2 → BR202012345
9. If it does not start with BR, search including BR and disregard the last 3 digits. Considering whether one or more words of the patent title informed on the Lattes Platform, removing stopwords, contain the patent title on Espacenet. If you find it, go to step 10, otherwise the patent is considered uncertified.
10. The stopwords and accents of the title of the patent informed in the curriculum are removed, as well as the title of each patent located. Subsequently, the title of the patent informed in the curriculum is compared with the title of each patent located, calculating the similarity of the titles using the Levenshtein method. Considering as certified only patents with title similarity greater than or equal to 85%.

Therefore, after all the steps described, it was possible to certify the patents that are in the curricula registered on the Lattes Platform, adding various other information, as well as having access to the web page for the registration of the said patent on Espacenet, which includes other information such as documents and images of the invention.

3   Results

As a result, through the strategies defined in this work, it was possible to collect 903,979 patent records from the Espacenet database. With regard to patent data from the Lattes Platform curricula, it was possible to collect 76,619 patent records distributed in 28,581 curricula (Tab. 1).

Table 1. Quantitative data collected

Record type	Amount
Patents collected on Espacenet	903.979
Patents collected on the Lattes Platform	76.619
Lattes Platform CVs with patent records	28.581

In turn, contemplating the main objective, it was possible to certify 44,348 of the total of 76,619 patent records collected in the Lattes Platform, a percentage of approximately 58% of the patents. Figure 1 presents a chart comparing certified and uncertified patents.
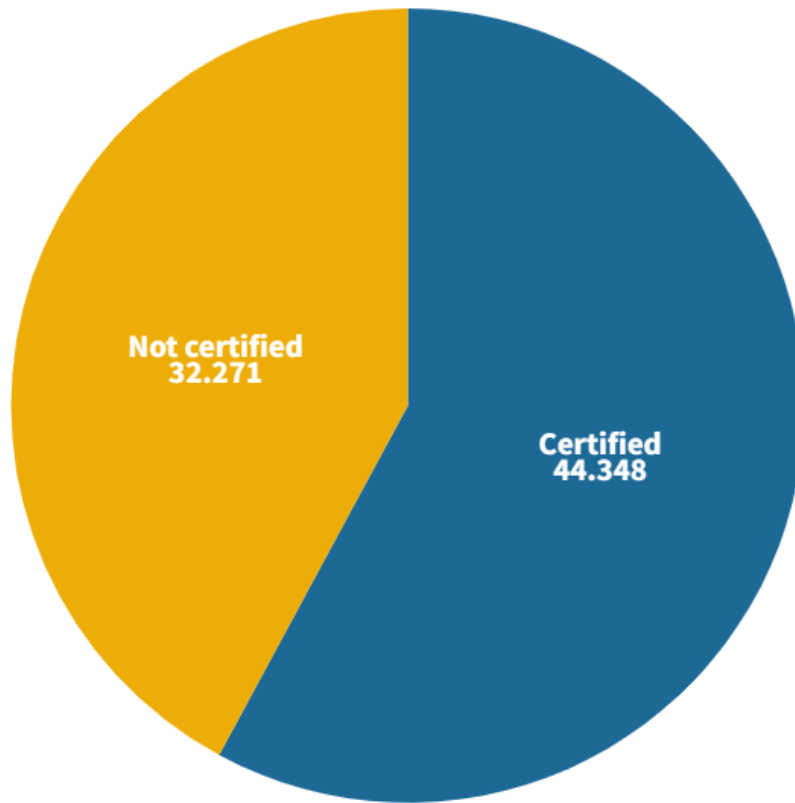
Figure 1. Certified patents and uncertified patents.

This result points to the need to have a patent data certifier informed on the Lattes Platform, given the high number of uncertified patents. With the certification of patents it is possible to increase the consistency of the data that are informed by the author himself.

With the aim of seeking to understand the reason for almost 42% of patent registrations not being certified, fig. 2 shows the total number of patents deposited per year, considering only the patents reported in the CVs, showing a series for certified ones and another series for patents. not certified.
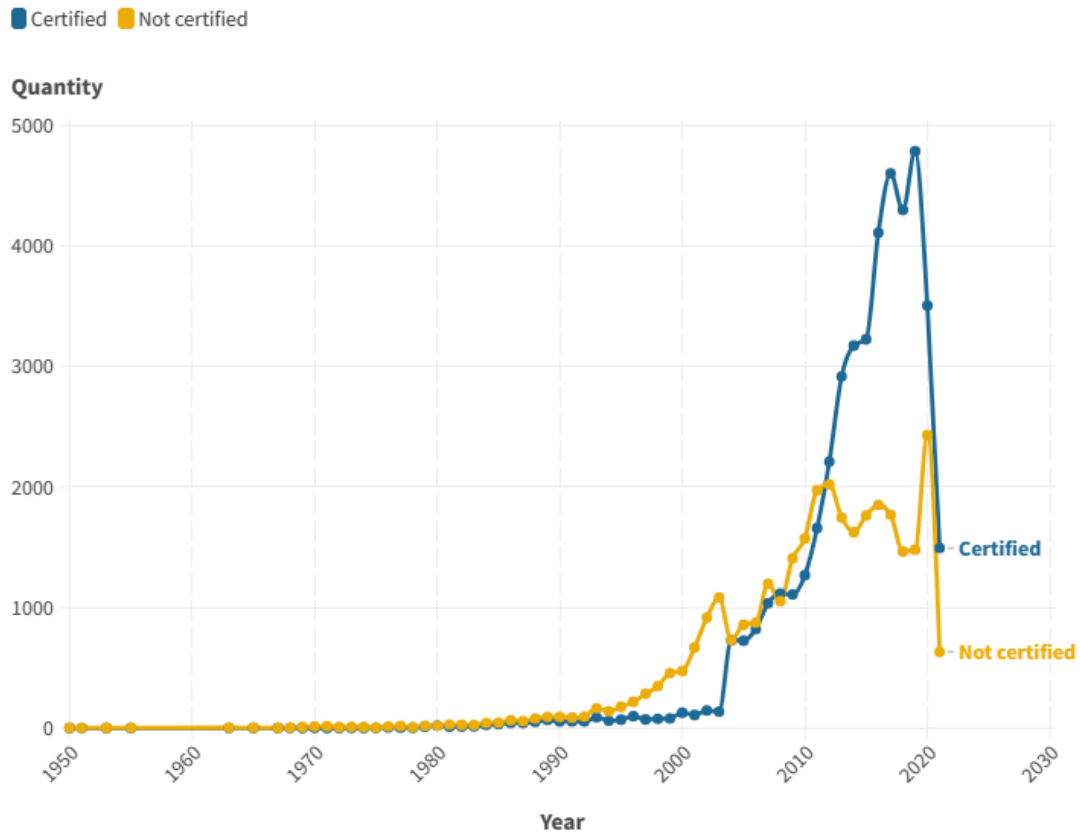
Figure 2. Timeline certified patents and non-certified patents.

It is possible to observe that between the years 1950 and 1990, there is not a considerable difference between the total number of certified and non-certified patents, however, between the years 1990 and 2002, the number of non-certified patents increased considerably. Considering the fact that the INPI has two different formats for defining the patent filing number, a first used for older patents and a second format for patents filed from January 2, 2012, aiming to meet the international patent identification standard, which it is even the standard adopted by Espacenet. Which implies that all patents filed until 2011 in Brazil, and which are available on Espacenet, had their numbering changed in order to meet the international standard (SILVA [8]). Based on what has been presented, it is possible to explain why between 1990 and 2002 the number of uncertified patents increased and after 2011 this scenario reversed, the number of certified patents exceeded the number of uncertified patents.

In the period from 1950 to 2011, 10,019 (39.30%) were certified and 15,476 (60.70%) were not certified. Between 2012 and 2021, 34,329 (67.16%) were certified and 16,785 (32.84%) were not certified. It is worth noting that if you add up all the values, you get a total of 76,609, thus missing 10 patent registrations from the number informed in Table 1, these 10 missing registrations were uncertified patents that did not have the year informed in the Lattes Platform curriculum, therefore, they were not included in the temporal analysis.

The strategy proposed in this work only certifies patents deposited in Brazil, therefore, it is important to check if there are patents that were not deposited in Brazil in the set of non-certified patents. Therefore, it was possible to identify that 7,977 uncertified patent records were not deposited in Brazil, justifying their non-certification through the defined strategies.

In continuity, after the identification of patents not deposited in Brazil, the set of non-certified patents now comprises 24,294 records, that is, about 32% of which present several inconsistencies that justify their non-identification, such as, for example, the filing number patents "62/1781"; "No.Ref. 05385"; "PI? 0006594-3"; "n.05059911-9" and "INPI PI0705367" which, after applying the strategy, did not generate numbers of deposits that could be certified nor by their titles. Another example was the patents with the filing number "WO2010054452", "US2009246833-A1", which do not meet the Brazilian patent numbering standard and were

reported by their proponents as Brazilian patents.

4   Conclusions

With the certification of the information inserted in the curricula, it is possible to validate that the scientific works, patents, software, among others, were inserted correctly, avoiding possible incorrect information. For this reason, the certification of patents reported in the Lattes Platform curricula stands out as a factor that promotes greater credibility and generates confidence in the veracity of the information. Furthermore, as shown by the results, the number of patents reported in the Lattes Platform curricula has been increasing over the years, it is believed that this percentage of curricula with patent information will increase.

Authorship statement.

The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] FERREIRA, Chrystiano de Campos; PIRES, Rafael Cardoso; FERREIRA, Herika Rangel; MEDEIROS, Fernanda Fancicani; ZAMBRANA, Diego Eusebio Quispe; LOPES, Wagner Bacelos; KORTE, Rodolfo Luis. How to write and publish a scientific article: a literature review. Revista JRG de Estudos Acadêmicos, Brasília, DF, v. 6, n. 12, p. 216-225, jan./jul. 2023.

[2] RIBEIRO, Mariana Eleutério; FREY, Irineu Afonso; AZEVEDO, Paola. Classification of Patents in Federal Universities on the TRL Scale (Technology Readiness Level): case study based on ISO 16290:2013. Cadernos de Prospecção, Salvador, BA, v. 15, n. 1, p. 117-130, jan./mar. 2022. DOI: https://doi.org/10.9771/cp.v15i1.42173.

[3] LOPES, Ricardo Cortez. Curriculum Lattes: a study on some social representations. PragMATIZES - Revista Latino-Americana de Estudos em Cultura, Niterói, RJ, v. 12, n. 22, p. 663-694, mar. 2022.

[4] DIGIAMPIETRI, Luciano Antonio; MENA-CHALCO, Jesus Pascual; PÉREZ-ALCAZAR, José de Jesús; TUESTA, Esteban Fernandez; DELGADO, Karina Valdivia; MUGNAINI, Rogério; SILVA, Gabriela S. Mining and Characterizing Data from Lattes Resumes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 1., 2012, Curitiba. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2012. ISSN 2595-6094.

[5] BRASIL. Ministério da Ciência, Tecnologia e Inovações. Conselho Nacional de Desenvolvimento Científico e Tecnológico. Oasisbr is now used for certification of Theses and Dissertations in the Lattes Curriculum: Lattes theses and dissertations now have the Oasisbr seal of authenticity certification. 12 set. 2022. Available in: http://bit.ly/3P9jjHm. Accessed on: 4 May 2023.

[6] ESPACENET. Espacenet patent search. 2023. Available in: https://worldwide.espacenet.com/patent/. Accessed on: 27 Apr. 2023.

[7] DIAS, Thiago Magela Rodrigues. A Study of Brazilian Scientific Production Based on Data from the Lattes Platform. 181 f. Thesis (Doctorate in Mathematical and Computational Modeling) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

[8] SILVA, Raulivan Rodrigo da. Collection and treatment of data on Brazilian technical production: a study based on patents. 2022. 77 f. Dissertation (Master in Mathematical and Computational Modeling) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2022.