XLIII CILAMCE
21–25, November, 2022
FOZ DO IGUAÇU – PR – BRAZIL

Collection and Processing of Data on Brazilian Technical Production in Engineering Areas

Raulivan R. Silva[1], Thiago M. R. Dias[1], Higor A. D. Mascarenhas[1]

[1]*Federal Center for Technological Education of Minas Gerais – CEFET-MG*
Av. Amazonas, 7675, 30510-000, Minas Gerais/Belo Horizonte, Brazil
raulivan@cefetmg.br, thiagomagela@cefetmg.br, higoralexandre1996@gmail.com

Abstract. The main objective of this article is to present a strategy for the identification and extraction of data from Brazilian patents from researchers working in the areas of Engineering, such as title, abstract, filing date, publication date, inventors, owners, among others. Such a strategy will allow the construction of a local database of the Brazilian technical production of individuals working in the engineering areas, enabling analysis of the large volume of data in a shorter time, since the analysis will be local and not in online repositories of patents. Therefore, the proposed solution will allow to minimize several restrictions imposed by online repositories, among them it is possible to mention the limit in the volume of data access and internet connectivity. In this study, the National Institute of Industrial Property (INPI) and the international patent repository Espacenet, of recognized international relevance, will be used as the main repositories. The obtained results allow us to verify how this type of production has evolved over the years, considering the technical production of individuals who work in the various areas of Engineering.

Keywords: Patents, INPI, Technical Production.

1 Introduction

Currently, several repositories available on the internet make it possible to search for published scientific productions, namely DBLP (Digital Bibliography Library Project), ArnetMiner, Google Scholar, Microsoft Academic Search and the CNPq Lattes Platform, the latter being an extremely data-rich instrument for studies on Brazilian scientific and technical production. Therefore, as with scientific productions, in the context of technical production, there are also repositories of patent records, such as the pePI (Industrial Property Research) maintained by the Brazilian patent management body INPI (National Institute of Intellectual Property). As in Brazil, each country has its own body responsible for managing the filing and granting of patents, as well as making them available for consultation. In addition, there are international repositories of patent registration, some of them, such as Espacenet, of recognized international relevance. Espacenet, which makes it possible to consult patents from approximately 70 countries, including Brazil, in a single repository, stands out in view of the amount of data available [1].

Although there are several patent consultation repositories, these have limitations when querying data, such as querying a large volume of data, limiting data traffic, limiting access via boot programming, allowing only human access, which makes the costly analysis [2].

Therefore, this work aims to present an alternative to enable the analysis of patents, especially in the areas of engineering. A strategy is proposed for the identification and extraction of data from Brazilian patents, enabling the construction of a local database of patents, which will allow the researcher to prepare their analyzes with greater freedom to manipulate, process and format the data according to their needs.

2 Methodology

This article is a case study, that is, an empirical study that investigates a particular phenomenon, within a context in which there are still gaps in the literature [3].

The collection of information regarding patent documents deposited at the INPI from 01/01/1900 to 12/31/2020 was carried out. Thus, in possession of the data, the patent data was consulted in the Espacenet repository, using the patent filing number collected at the INPI. This set of data extracted at the INPI as well as at Espacenet is the set of data that will compose the database, seeking to guarantee the consistency of the data collected.

A. Data acquisition

The patent data acquisition process was divided into two stages, (1) initially the data collection at the INPI and processing of the patent filing numbers, and later, (2) the validation and collection of patent data from the patents extracted at the Espacenet repository. Figure 1 presents the scheme developed for the data collection process.

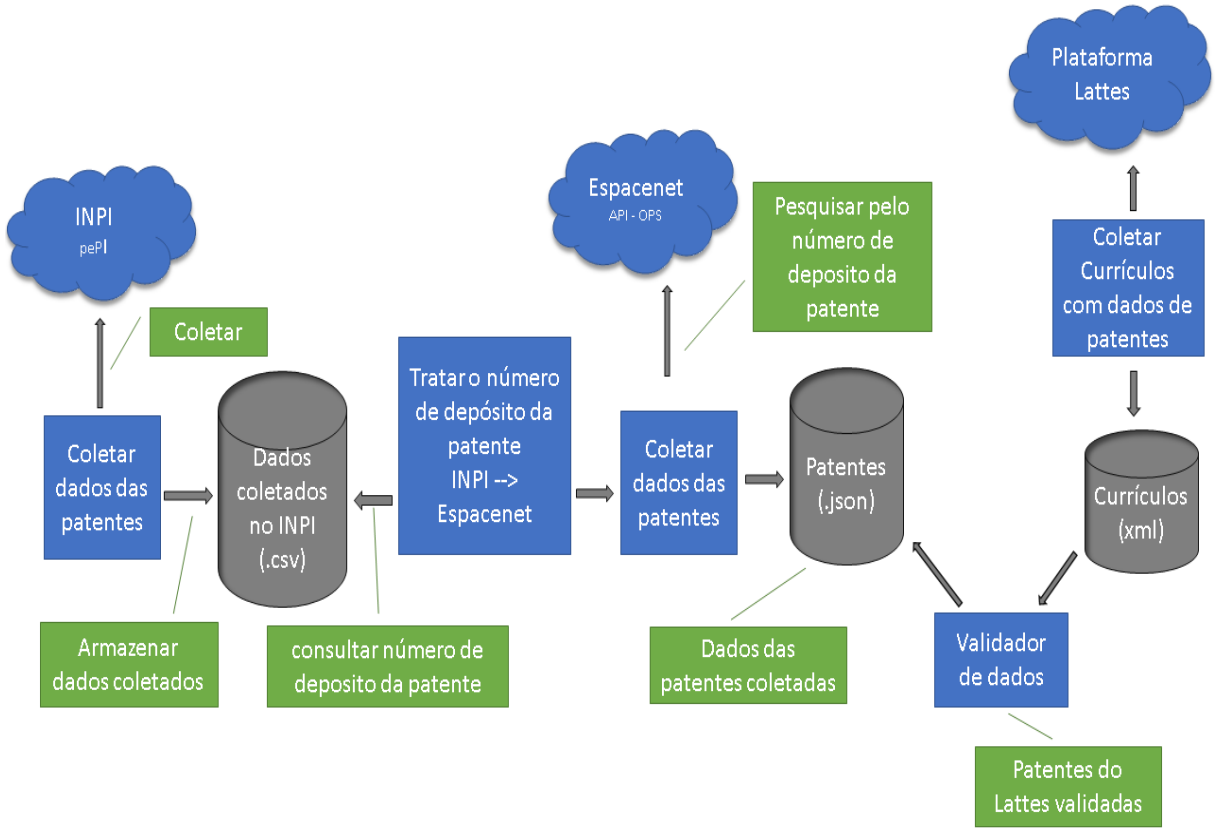

Figure 1. Overview of data collection

1) Data collection at the INPI: To collect the patent data at the INPI, the patent search tool pePI (Industrial Property Research) maintained by the INPI was used, where it is possible to consult patent documents by informing login and password or by anonymous access. What differs between the two forms of identification is that choosing to inform the login and password will allow access to more services, such as the availability of documents in PDF format, among others, but to achieve the objective of this work, anonymous access is the enough, therefore, the same was used. After accessing the tool using the anonymous identification method, the option “Patents” was selected, where a page containing search options is presented, the option “advanced search” was selected to display more search criteria.

When entering in the search field "(22) Deposit Date" the start date "01/01/1900" and the end date "12/31/2020" and select the option "search", the system returns a page with the list of 862,726 patents spread over 8,627 pages displaying 100 records per page. As an illustration, if we performed the manual collection of all patents, this would require a very large human effort, considering that it would take approximately 10 minutes to access the details of each patent and store the information of interest in a spreadsheet processor, it would take about 143,434 hours, devoting 8 hours a day would take about 17,942 days. To optimize data collection, an algorithm was proposed to enable a computational process in order to automate the collection, consisting of 5 steps:

1. Perform anonymous login to retrieve the credentials needed to perform the survey;
2. Access the advanced search, informing the credentials obtained in the previous step;
3. On the advanced search screen, enter the start date 01/01/1900 and the end date 12/31/2020 in the field "(22) Deposit Date" and trigger the search event;
4. Scroll through the entire patent listing displayed on the result page:
    A. The. For each patent, access the details page;
    B. Analyze the HTML (HyperText Markup Language) content of the detail page and retrieve the information: "Order number", "Deposit date", "Publication date", "Title", "Depositor", "Inventor" and "Classification" ICP".
    C. Store the retrieved information in a CSV (Comma-separated-values) file;
    D. Back to patent listing;
5. Repeat step 4 for all search results pages.

Through web scraping and web crawler techniques, this entire strategy was coded using the Python programming language. Zhao [4] defines web scraping, which in translation into Portuguese, web scraping, as a technique for extracting data from pages available on the WWW (World Wide Web) and storing it in a file or database for further analysis. be performed manually by a user or automatically by a robot (web crawler). Web crawler is an algorithm used to find, read and index pages on a website. Web scraping encompasses a large set of programming techniques and several technologies, such as data analysis, parsing of natural languages and information security, among others [5].

During the tests of the developed algorithm, it was possible to identify a limitation in the established approach, due to the large volume of data, for reasons of platform security, the credentials expire after a certain time. To circumvent this limitation, monthly periods were used for the "Deposit Date" filter, thus achieving, in only 0.5% of the time, when compared to the manual process, to collect patent information. Therefore, storing the data in CSV files, one file for each year, the collection was performed between the months of April and June 2020.

2) Data collection at Espacenet: With the data collection at the INPI completed, the next step was to identify each patent collected at the INPI, at Espacenet, and later extract its available data. Only the set of patents that are identified on Espacenet will be considered, due to the completeness and consistency of the data.

Espacenet is a worldwide intelligent search service that provides free access to information on inventions and technical developments from 1782 to the present day. Its query interface is simple and intuitive, making it accessible even to inexperienced users, currently containing data from more than 120 million patent documents from all over the world [2].

The Platform offers smart search features, in which it is possible to enter the desired term where it is searched in several fields of the patent, being able to enter up to 10 terms separated by space. The service was designed to be used by humans, not allowing automatic queries or batch retrieval, when this is necessary, the use of OPS (Open Patent Services) is recommended. OPS is a web service that provides access to data stored in the EPO (European Patent Office) database through web services using RESTful architecture. Making use of XML (eXtensible Markup Language) and JSON (JavaScript Object Notation) standards to format the response data to requests, according to the parameterization. This consequently makes it viable to develop applications and self-extracting robots to download large volumes of data.

The retrieval of data referring to each patent is made possible using the patent search available at OPS, using the patent application number as a selection criterion. The order number is important for patent identification both at INPI and Espacenet, as each patent has its own unique filing number. The composition of the patent filing application number at the INPI has two different formats, the first used for older patents and a second format is currently adopted. Since January 2, 2012, new applications for patents (invention and utility model), industrial design and geographical indication are assigned the new format [6].

The format assigned to patents filed up to 12/31/2011 is composed of the following format ZZ XXXXXXX-D, where ZZ refers to the nature of the protection, XXXXXXX an annual serial number composed of 7 digits, and finally, D, which is the verifying digit. Figure 2 graphically presents the composition of the format.

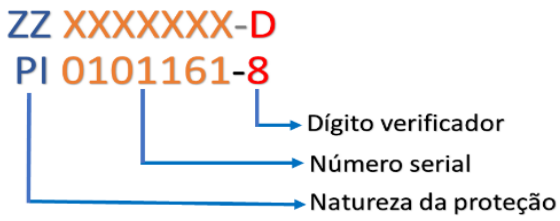

Figure 2. Old patent numbering format

The new format established aims to meet the INPI's international integration policy, meeting the standards internationally suggested by the World Intellectual Property Organization (WIPO). This new format has the following structure BR ZZ AAAA XXXXXX D CP, where BR is the identification of the country, ZZ is the nature of the protection, YYYY year of entry into the INPI, XXXXXX numbering that corresponds to the order of filing of applications composed of 6 digits, D, the check digit and finally CP that corresponds to the publication code, the legal status of the request with the INPI. Figure 3 graphically presents the composition of the current numbering format for patent applications.

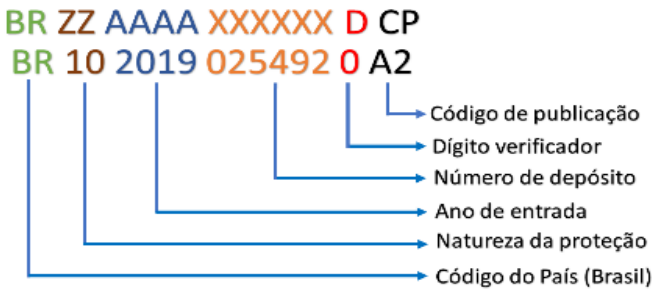

Figure 3. New numbering format

After processing the patent filing numbers, an algorithm was developed using the Python programming language, which runs through all the CSV files with the results of processing the patent filing numbers and making use of the services available in the OPS, performs the consultation of each patent, using the previously processed patent application numbers as search criteria, storing each patent located in the Espacenet repository, in a file in .json format.

As a result, 722,347 patents were identified in the Espacenet repository, about 83% of the set of patents collected at the INPI. A hypothesis for unidentified patents is due to the fact that they have not yet been made available in the Espacenet repository, or due to problems in identifying the correct format of the patent application number, which may be identified and addressed in future works.

With the success of collecting the patents, it was also possible to collect, through Espacenet, the families of Brazilian patents, so far 21% of the families have been collected and stored in JSON files.

All data collected add up to a total of 23.7 GB of data on Brazilian technical production.

3 Patents in the areas of Engineering

Continuing the data collection, the next step was to collect CVs registered on the Lattes Platform that have patent information, such as the filing application number or the patent title, in order to validate such information with the set collected from Espacenet. Such validation is important because all curricular information on the Lattes Platform is included by the individual, and this validation process and later certification of the data is important. Lattes Platform curricula are freely available on the internet for consultation, recording all their professional, academic and scientific productions. The process of collecting and selecting curriculum data from the Lattes Platform was carried out using the LattesDataXplorer framework. The framework has a set of techniques and methods responsible for collecting, selecting, processing and analyzing data (Figure 4).

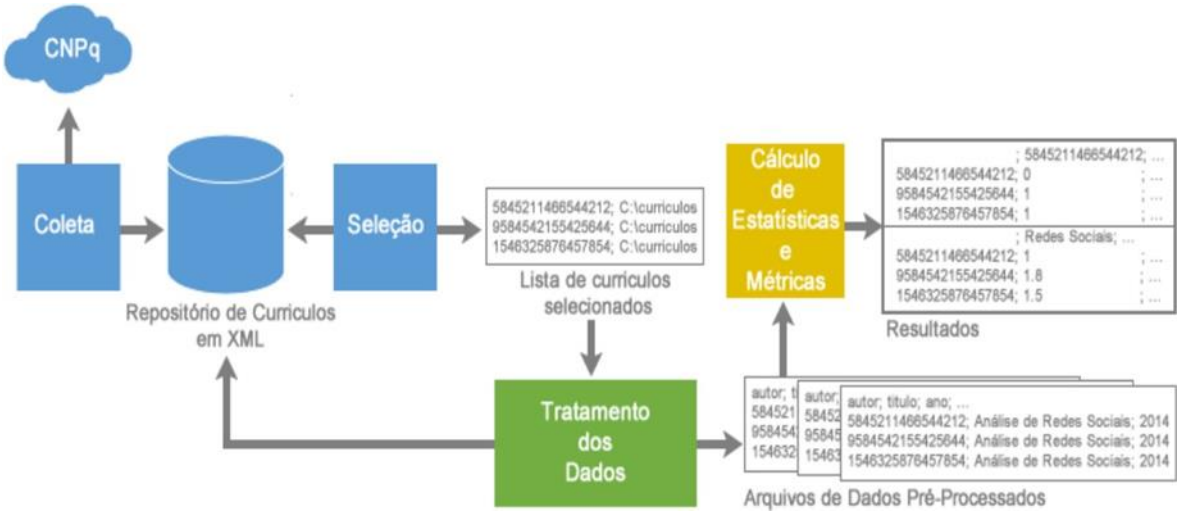

Figure 4. LattesDataXplorer overview

The LattesDataXplorer framework has a rich framework composed of several modules, to extract and select the curriculum data from the Lattes Platform, the collection and selection modules were used. The entire collection process consists of 3 steps:

1. URL extraction: responsible for obtaining the unique codes of all registered curricula, allowing individual access to each curriculum, as each curriculum in the Lattes Platform has a unique code composed of eight digits.
2. Extraction of Identifiers and Date: Access the header of each curriculum to retrieve its unique identifier, the last update date and the curriculum code. Storing the data locally in an identification file that contains code, identifier, date of last update on CNPq and date of curriculum update.
3. Extraction of CVs: Extract and store CVs. If the update date on the Lattes Platform differs from the update date of the locally stored curriculum, the local curriculum is replaced by the most recent one. The CVs collected are in XML format, as this version contains well-delimited sections and fields.

The collection of CVs was carried out in July 2020, retrieving 72,256 records with patent information distributed in a total of 29,516 CVs. The extractor collects the CVs and stores them in XML format in folders, identified from 00 to 99. The name of the selected folder to store the file and the file name are defined according to its unique 16-digit identifier number, the two first numbers of the identifier correspond to the name of the folder and the remaining 14 digits correspond to the name of the saved file.

In the Lattes Platform curricula, individuals can inform their main areas of activity in their data. This fact makes it possible to ascertain which area of knowledge is most representative in view of the profile of the analyzed proponents, and consequently, to assess which area of knowledge is characterized as a propeller of national technological development. Figure 5 presents the major areas of knowledge and their respective numbers of individuals. The selection criterion adopted was to count all the CVs that have at least one valid patent on Espacenet and that have information on the area of expertise registered in their CV. As a curriculum can have up to six records of areas of activity, for this analysis the first area informed was considered.

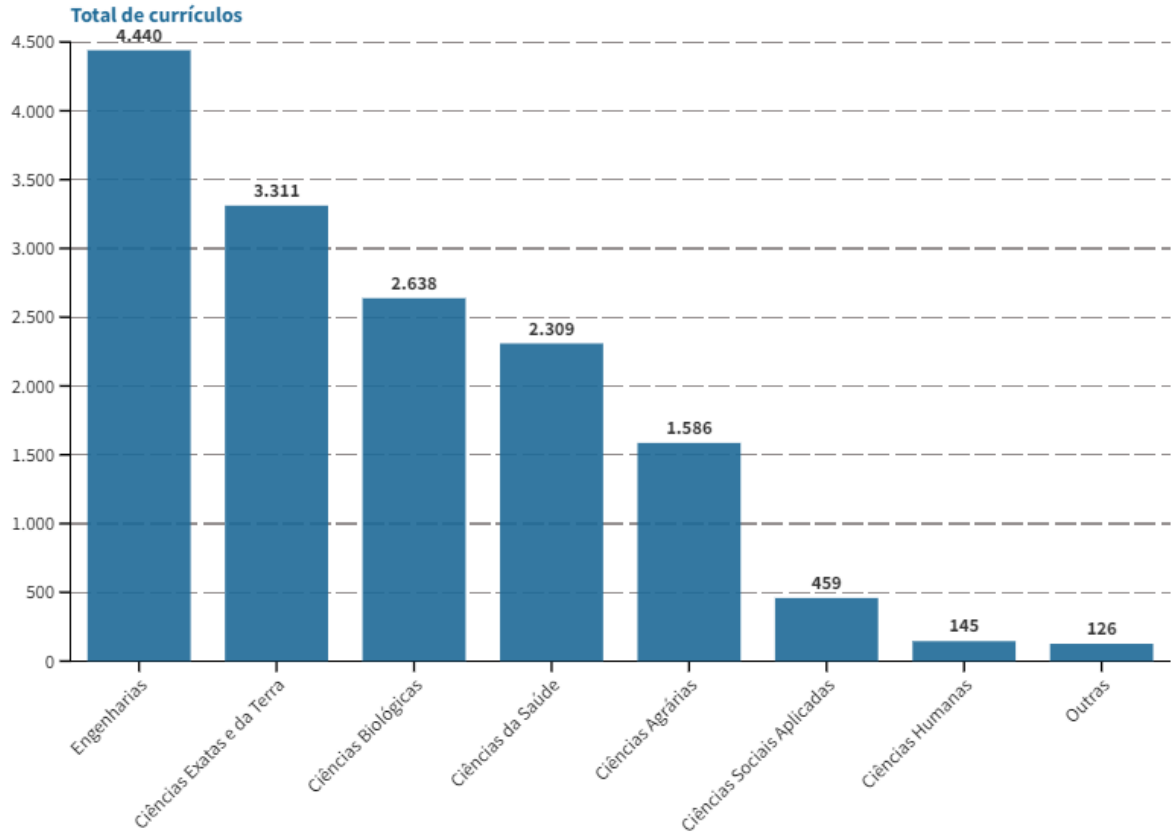

Figure 5. Large area of activity of patent applicants

Technological development is directly linked to the most diverse areas of knowledge, which implies the constitution of multidisciplinary teams with a high level of knowledge, however, the area of engineering stands out in the technological development in the analyzed set.

In this context, it is worth mentioning that in possession of this set of patents deposited by individuals whose main area of activity is Engineering, different metrics and analyzes can be applied. The set of patents in the engineering areas, as can be seen, is characterized as the largest set holding all those identified, thus corroborating as an important source for analyzes of Brazilian technical production.

4 Conclusions

Technological innovation and the competitiveness of countries and companies in a globalized scenario aimed at sharing information and knowledge are directly associated with Intellectual Property rights. Companies and Universities feel a growing need to ensure legal protection for products and/or technologies they develop, contributing to the continuous and growing number of patents being filed, thus constituting a set of data with

valuable information on technological development. Through the analysis of patent documents, it is possible to investigate: which patents were most deposited by a given organization in a time window or throughout its trajectory; identify emerging technologies; which organizations or individuals are patenting in a particular area.

Due to the results obtained, it was possible to characterize the Brazilian technical production, distributed by the various areas of activity of its proponents, with emphasis on individuals working in the areas of Engineering. It was possible to highlight the importance of adopting information contained in patent documents as a source of data for analysis of technical production. Although quantitative indicators are not enough to show the value of a patent, it allows us to understand the entire ecosystem of intellectual property protection. The indicators contribute significantly to the monitoring of technological development, serving as a reference for prospecting new technologies. Therefore, it is possible to affirm that patentometry has a great scientific value and needs to be further explored in the national scenario.

Acknowledgements. The authors thank the Federal Center for Technological Education of Minas Gerais (CEFET/MG) for their research assistance.

Authorship statement. This section is mandatory and should be positioned immediately before the References section. The text should be exactly as follows: 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] ESPECENET. Espacenet patent search. 2021. Available in: <https://worldwide.espacenet.com/patent/>.

[2] BRANDÃO, F. G. Democratização da informação a partir do uso de repositórios digitais institucionais: da comunicação científica às informações tecnológicas de patentes. Dissertação (Mestrado) - Universidade Regional Integrada do Alto Uruguai e das Missões, sep 2016. Available in: <https://lume.ufrgs.br/handle/10183/179853>.

[3] SERRANO, B. P.; JUNIOR, J. A. G. Redes de inovação: mapeamento de inventores de patentes em uma empresa do setor de cosméticos. Revista GEPROS, v. 09, n. 1, p. 101,jan 2014.

[4] ZHAO, B. Web scraping.Springer International Publishing, p. 1–3, may 2017. Available in: <https://www.researchgate.net/publication/317177787_Web_Scraping>.

[5] MITCHELL, R.Web Scraping com Python: Coletando mais dados da web moderna. second. [S.l.]: Novatec Editora., 2019.

[6] UECE, U. F. do C.INPI - Saiba mais sobre a nova numeração nos pedidos da DIRPA e da DICIG. 2011. Available in:<http://www.uece.br/nit/index.php?option=com_content&view=article&id=1654:inpi-saiba-mais-sobre-a-nova-numeracao-nos-pedidos-da-dirpa-e-da-dicig&catid=31:lista-de-noticias>.