XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT-8 – Informação e Tecnologia

UMA ESTRATÉGIA PARA RECUPERAÇÃO DE DADOS PARA ANÁLISES SOBRE A PRODUÇÃO TÉCNICA BRASILEIRA

A STRATEGY FOR DATA RECOVERY FOR ANALYSIS ON BRAZILIAN TECHNICAL PRODUCTION

Raulivan Rodrigo da Silva - Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Thiago Magela Rodrigues Dias - Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Washington Luís Ribeiro de Carvalho Segundo - Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

Modalidade: Resumo Expandido

Resumo: Este trabalho tem como principal objetivo apresentar uma visão geral da produção técnica brasileira com base na análise de patentes registradas no Instituto Nacional da Propriedade Industrial (INPI), bem como, de repositórios internacionais como por exemplo a Espacenet. Inicialmente, utilizando a base curricular da Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), o presente estudo busca efetuar uma análise das patentes registradas nos currículos dos pesquisadores brasileiros e sua validação junto ao repositório da Espacenet, no intuito de verificar a validade dos dados, bem como obter um conjunto de informações complementares sobre as patentes recuperadas.

Palavras-Chave: patentes; INPI; plataforma Lattes; espacenet; patentometria.

Abstract: The main objective of this work is to present an overview of Brazilian technical production based on the analysis of patents registered at the National Institute of Industrial Property (INPI), as well as international repositories such as Espacenet. Initially, using the curricular base of the Lattes Platform of the National Council for Scientific and Technological Development (CNPq), this study seeks to carry out an analysis of patents registered in the curricula of Brazilian researchers and their validation with the Espacenet repository, in order to verify the validity of the data, as well as obtaining a set of additional information about the recovered patents.

Keywords: patents; INPI; Lattes platform; espacenet; patentometry.

1 INTRODUÇÃO

O século XXI tem sido solo fértil para a criação de estruturas tecnológicas, e mais do que nunca, a rapidez na evolução destas tecnologias tem sido visível. Diariamente novos dispositivos, aplicações, meios digitais permeiam o mercado, trazendo versões melhores de

recursos e/ou funcionalidades que até então conhecíamos ou apresentando novas soluções. Em consequência disso, as organizações desse mercado estão empenhadas em realizar um monitoramento constante de suas atividades e da viabilidade de seus produtos e serviços oferecidos, sendo necessário para tanto implementar inovações que fidelizem ou aumentem a base de clientes (AMADEI; TORKOMIAN, 2009). Boa parte das invenções e inovações geradas possuem como origem pesquisas iniciadas em instituições de ensino públicas e privadas, que vem aumentando ao longo dos anos tornando-as um grande polo de inovação nacional. Este fator impulsionou a criação de leis e normas que visam proteger a propriedade intelectual gerada em universidades e também núcleos que visam auxiliar no processo de criação de patentes e viabilizando o relacionamento entre os setores acadêmicos e mercadológicos, fomentando para que instituições de ensino e pesquisadores se inclinem cada vez mais ao depósito de patentes (VALMIRA; PERUCCHI, 2014).

É possível encontrar diversas pesquisas que se baseiam no número de patentes depositadas e no número de patentes concedidas, usando-as como parâmetros para determinar o volume de inovação tecnológica de um determinado país ou instituição. Contudo, isso não é o suficiente para compreender todo o cenário. Conforme Cativelli (2020) afirma, não é possível afirmar que todas as patentes realmente contribuem com o crescimento da ciência e tecnologia, ou até mesmo se geram frutos para seus titulares e inventores.

Logo, assim como ocorre com as produções científicas, no contexto da produção técnica, existem também repositórios de registros de patentes, como o pePI (Pesquisa em Propriedade Industrial) mantido pelo órgão brasileiro de gestão de patentes INPI (Instituto Nacional de Propriedade Intelectual). Assim como no Brasil, cada país possui seu órgão responsável por gerenciar o depósito e concessão de patentes bem como disponibilizá-las para consulta. Além disso, existem repositórios internacionais de registro de patentes, sendo alguns deles como a Espacenet de reconhecida relevância. A Espacenet, que viabiliza consultar em um único repositório patentes de aproximadamente 70 países, se destaca tendo em vista a quantidade de dados disponibilizada.

Diante disso, este trabalho visa ampliar a compreensão sobre as atividades de patenteamento construídas nacionalmente, buscando avaliar os principais atores, as redes de colaboração ocultas e o resultado que estas infligem na evolução da ciência. Para tanto, este estudo propõem um ferramental tecnológico para a coleta e tratamentos de dados das

patentes nacionais disponibilizadas no repositório da Espacenet, visando construir um repositório local que viabilize todas as análises objetivadas.

2 METODOLOGIA

Este artigo trata-se de um estudo de caso, ou seja, um estudo de natureza empírica que investiga um determinado fenômeno, dentro de um contexto em que ainda há lacunas na literatura, conforme afirmam (SERRANO; JUNIOR, 2014).

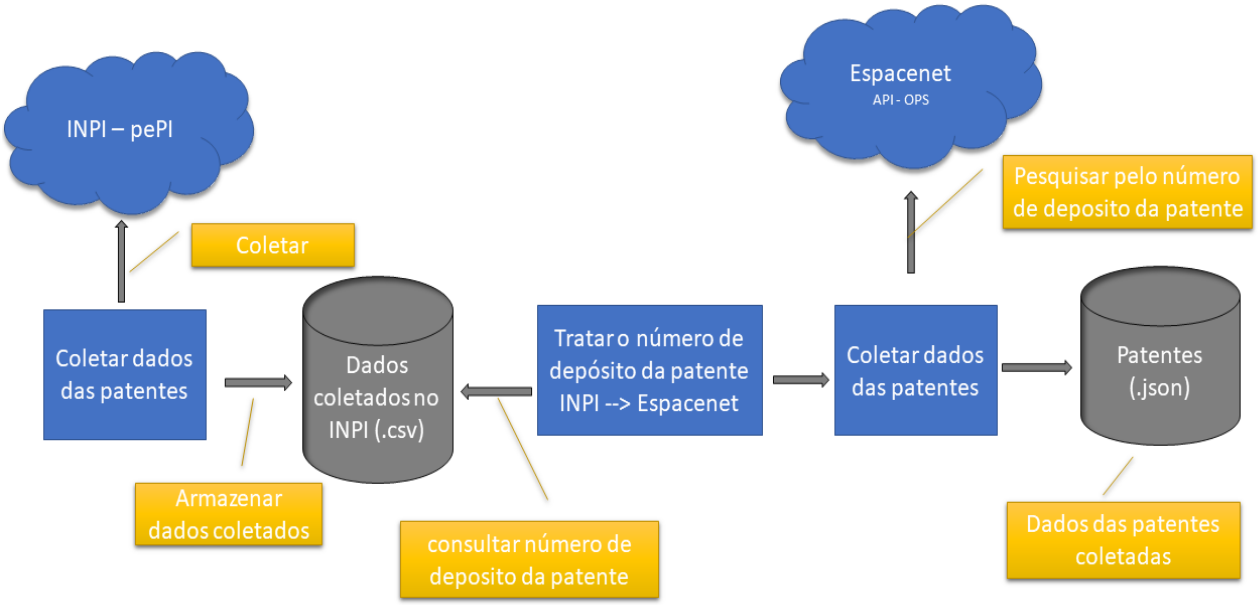
A priori foi realizada a coleta de informações referente a documentos de patentes depositadas no INPI no período de 01/01/1900 à 31/12/2020. De posse dos dados, foi realizada a consulta de dados patentários no repositório da Espacenet, utilizando para isso o número de depósito das patentes coletadas no INPI. Esse conjunto de dados extraídos no INPI como também na Espacenet é o conjunto de dados analisado neste estudo, estudo este de grande relevância, principalmente pelas técnicas implementadas, bem como, amplitude e consistência dos dados coletados.

Após a coleta dos dados das patentes, utilizando o framework *LattesDataXplorer* desenvolvida por Dias (2016), realizou-se a coleta de dados contidos nos currículos cadastrados na Plataforma Lattes do CNPq, que possuem informações de registro e/ou participação no depósito de patentes. De posse desses dados, algoritmos foram desenvolvidos para equalizar toda a base de dados obtida dos currículos, e também, para remover informações de usuários sem produção técnica registrada ou sem número válido de registros das patentes. O objetivo desta estratégia é verificar a consistência dos dados de patentes registrados nos currículos cadastrados na Plataforma Lattes, bem como, viabilizar análises que possam considerar diversas informações dos proponentes que não estão nos registros das patentes, mas informadas nos seus currículos, como por exemplo, dados de formação acadêmica, áreas de atuação e produção científica.

2.1 Aquisição dos Dados

O processo de aquisição dos dados das patentes foi dividido em duas etapas, (1) inicialmente a coleta dos dados no INPI e tratamento dos números de depósito de patentes, e posteriormente, (2) a validação e coleta de dados patentários das patentes extraídas no repositório da Espacenet. A Figura 1 apresenta o esquema elaborado para o processo de coleta dos dados.

Figura 1 – Visão geral da coleta de dados



Fonte: Elaboração do autor.

Para coletar os dados de patentes no INPI foi utilizada a ferramenta de pesquisa de patentes pePI (Pesquisa em Propriedade Industrial) mantido pelo INPI, onde é possível realizar a consulta de documentos de patentes informando login e senha ou por acesso anônimo. O que difere as duas formas de identificação é que optando por informar o login e senha irá permitir acessos a mais serviços, como por exemplo, a disponibilização de documentos no formato PDF entre outros, porém para atingir o objetivo deste trabalho o acesso anônimo é o suficiente, por isso, o mesmo foi utilizado.

Ao informar no campo de pesquisa “(22) Data de Depósito” à data inicial “01/01/1900” e data final “31/12/2020” e selecionar a opção “pesquisar”, o sistema retorna uma página com a listagem de 862.726 patentes distribuídas em 8.627 páginas exibindo 100 registros por página. Para otimizar a coleta dos dados, foi proposto um algoritmo para viabilizar um processo computacional no intuito de automatizar a coleta, composto por 5 etapas:

1. Realizar o login anônimo para recuperar as credenciais necessárias para realizar a pesquisa;
2. Acessar a pesquisa avançada, informando as credenciais obtidas na etapa anterior;
3. Na tela de pesquisa avançada, informar no campo “(22) Data de Depósito” a data inicial 01/01/1900 e a data final 31/12/2020 e disparar o evento de pesquisa;

4. Percorrer a toda a listagem de patentes apresentada na página de resultado
    a. Para cada patente, acessar a página de detalhes;
        i. Analisar o conteúdo HTML (*HyperText Markup Language*) da página de detalhe e recuperar a informações: “Número do pedido”, “Data de depósito”, “Data de publicação”, “Título Depositante”, “Inventor” e “Classificação ICP”.
        ii. Armazenar as informações recuperadas em um arquivo CSV (*Comma-separated-values*);
        iii. Voltar a listagem de patentes;
2. Repetir a etapa 4 para todas as páginas de resultados da pesquisa.

Por meio de técnicas de *web scraping* e *web crawler*, toda essa estratégia foi codificada utilizando a linguagem de programação Python.

Durante os testes do algoritmo desenvolvido foi possível identificar uma limitação nessa abordagem, devido ao grande volume de dados, por motivos de segurança da plataforma, as credenciais expiram depois de um determinado tempo. Para contornar essa limitação, foi utilizado períodos mensais para o filtro “Data de Deposito”. Logo, armazenando os dados em arquivos CSV, um arquivo para cada ano, a coleta foi executada entre os meses de abril a junho de 2020.

Com a coleta de dados no INPI concluída, a próxima etapa foi identificar cada patente coletada no INPI, na Espacenet, e posteriormente extrair seus dados disponibilizados. Somente o conjunto de patentes que forem identificadas na Espacenet será considerado, devida sua completude e consistência dos dados.

A Espacenet é um serviço de pesquisa inteligente de cobertura mundial que oferece acesso gratuito a informações sobre invenções e desenvolvimentos técnicos desde os anos de 1782 até a atualidade. Sua interface de consulta é simples e intuitiva, tornando-a acessível mesmo para usuários inexperientes, contendo atualmente dados de mais de 120 milhões de documentos de patentes de todo o mundo (ESPECENET, 2021). A Plataforma oferece recursos de pesquisa inteligente, em que é possível informar o termo desejado onde este é pesquisado em diversos campos da patente, podendo informar até 10 termos separados por espaço. O serviço foi projetado para ser usado por seres humanos, não permitindo realizar consultas

automáticas ou recuperação em lotes, quando isso é necessário é recomendado o uso do OPS (*Open Patent Services*).

OPS é um serviço da web que fornece acesso aos dados armazenados no banco de dados do EPO (*European Patent Office*) por meio de serviços web usando a arquitetura *RESTful*. Fazendo uso dos padrões XML (*eXtensible Markup Language*) e JSON (*JavaScript Object Notation*) para formatar os dados de resposta às requisições, conforme a parametrização. Consequentemente, torna-se viável o desenvolvimento de aplicativos e robôs de extração automática para baixar grandes volumes de dados.

A recuperação dos dados referente a cada patente é viabilizada utilizando a pesquisa de patentes disponível na OPS, usando o número de pedido de depósito da patente como critério de seleção. O número do pedido é importante para identificação da patente tanto no INPI quanto na Espacenet, pois cada patente possui seu próprio número único de depósito. A composição do número de pedido de depósito das patentes no INPI tem dois formatos distintos, em que um foi utilizado para patentes mais antigas e atualmente é adotado um outro formato. Desde 02 de janeiro de 2012 para os novos pedidos de patente (de invenção e modelo de utilidade), desenho industrial e indicação geográfica é atribuído o novo formato (UECE, 2011).

O formato atribuído às patentes depositadas até de 31/12/2011 é composto pelo seguinte formato ZZ XXXXXXX-D, onde ZZ é referente a natureza da proteção, XXXXXXX um número serial anual composto por 7 dígitos, e por fim, D que é o dígito verificador.

O novo formato estabelecido visa atender à política de integração internacional do INPI atendendo os padrões sugeridos internacionalmente pela OMPI St133 publicado pela WIPO (*World Intellectual Property Organization*). Esse novo formato possui a seguinte estrutura BR ZZ AAAA XXXXXX D CP, em que BR é a identificação do país, ZZ é a natureza da proteção, AAAA ano de entrada no INPI, XXXXXX numeração que corresponde a ordem de depósito dos pedidos composto por 6 dígitos, D o dígito verificador e por fim CP que corresponde ao código de publicação, o status legal do pedido junto ao INPI.

A Espacenet adota o padrão internacional OMPI St13 para armazenar as informações sobre as patentes, isso implica em realizar tratamentos no número de pedido de depósito antes de realizar a busca na Espacenet.

Tomando como base um conjunto de regras definidas, foi desenvolvido utilizando a linguagem de programação Python, um algoritmo que percorre todas as patentes coletadas no

INPI, e aplica todas as regras definidas armazenando os resultados em arquivos CSV, um arquivo para cada ano de depósito.

Após o tratamento dos números de pedido de depósito das patentes, foi desenvolvido um algoritmo utilizando a linguagem de programação Python, que percorre todos os arquivos CSV com os resultados do tratamento dos números de depósito de patentes e fazendo uso dos serviços disponíveis na OPS, realiza a consulta de cada patente, usando como critério de busca os números de pedido de depósito de patente previamente tratados, armazenando cada patente localizada no repositório da Espacenet, em um arquivo no formato .json. Após 241 horas de execução do algoritmo foi possível recuperar dados de 722.347 patentes identificadas com sucesso na Espacenet, cerca de 83% das 862.726 patentes coletadas no INPI. A coleta foi realizada entre os meses de julho a dezembro de 2020 e janeiro e fevereiro de 2021.

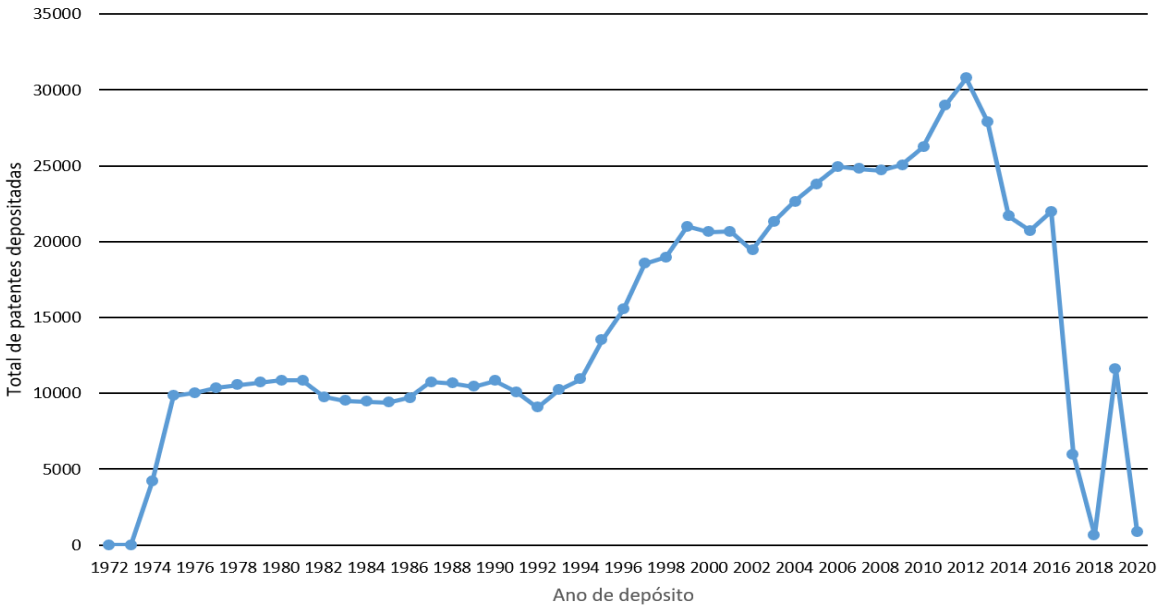
A próxima etapa deste trabalho foi coletar currículos registrados na Plataforma Lattes que possuem informações de patentes, como o número do pedido de depósito ou o título da patente, visando validar tais informações com o conjunto coletado junto a Espacenet. O processo de coleta e seleção dos dados curriculares da Plataforma Lattes foi realizado por meio do framework LattesDataXplorer (Dias, 2016).

3 RESULTADOS

Como resultado, inicialmente, 722.347 patentes foram identificadas no repositório da Espacenet, cerca de 83% do conjunto de patentes coletado no INPI, uma hipótese para as patentes não identificadas se dá pelo fato, que ainda não foram disponibilizadas no repositório da Espacenet, ou por problemas em identificar o formato correto do número do pedido de depósito da patente.

De posse do conjunto de dados de patentes recuperados no repositório da Espacenet, foi realizada a análise dos depósitos anual de patentes, apresentando dados entre 1972 a 2020, sendo o ano com maior número de patentes depositadas o ano de 2012, com um total de 30.774 pedidos de depósito (Figura 2). Destaca-se ainda que houve um crescimento contínuo no número de depósitos até o ano de 2012. E posteriormente, uma queda significativa.

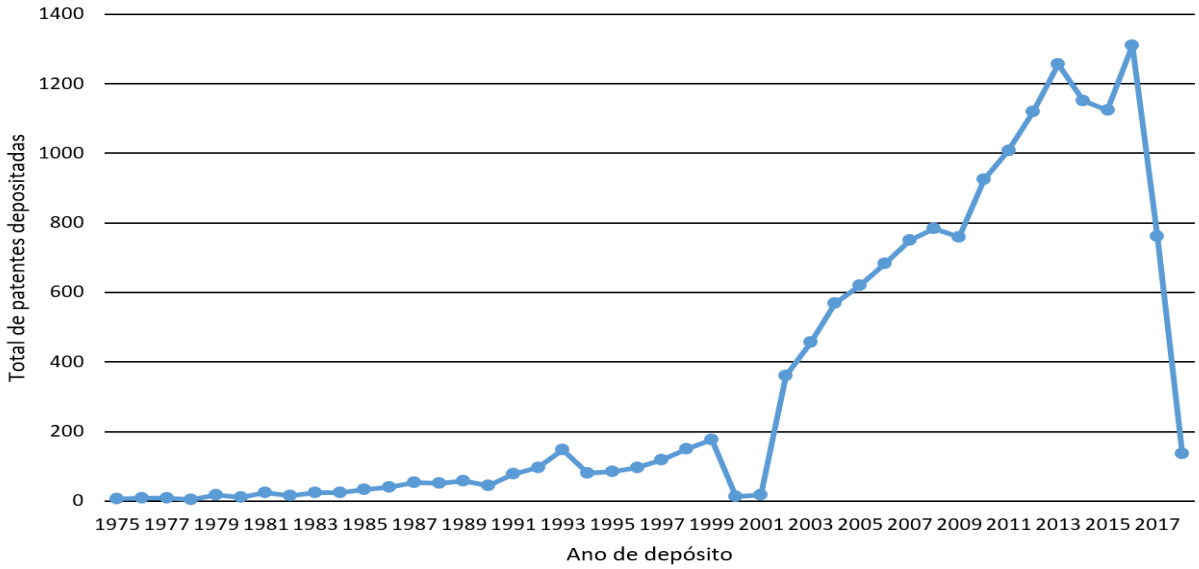
Figura 2 – Evolução temporal do deposito de patentes por ano



Fonte: Elaboração do autor.

No intuito de avaliar a representatividade dos registros de patentes cadastros nos currículos da Plataforma Lattes e devidamente identificadas na Espacenet, uma análise utilizando uma verificação entre os conjuntos foi realizada. Do total de 72.256 registros com dados de patentes extraída de 29.514 currículos, após o tratamento e limpeza dos dados referente a patentes extraídas dos currículos da Plataforma Lattes, foi possível identificar 15.252 patentes do conjunto extraído na Espacenet, a Figura 3 apresenta a distribuição das patentes identificadas nos currículos por ano de depósito.

Figura 3 – Deposito de patentes nos currículos da Plataforma Lattes

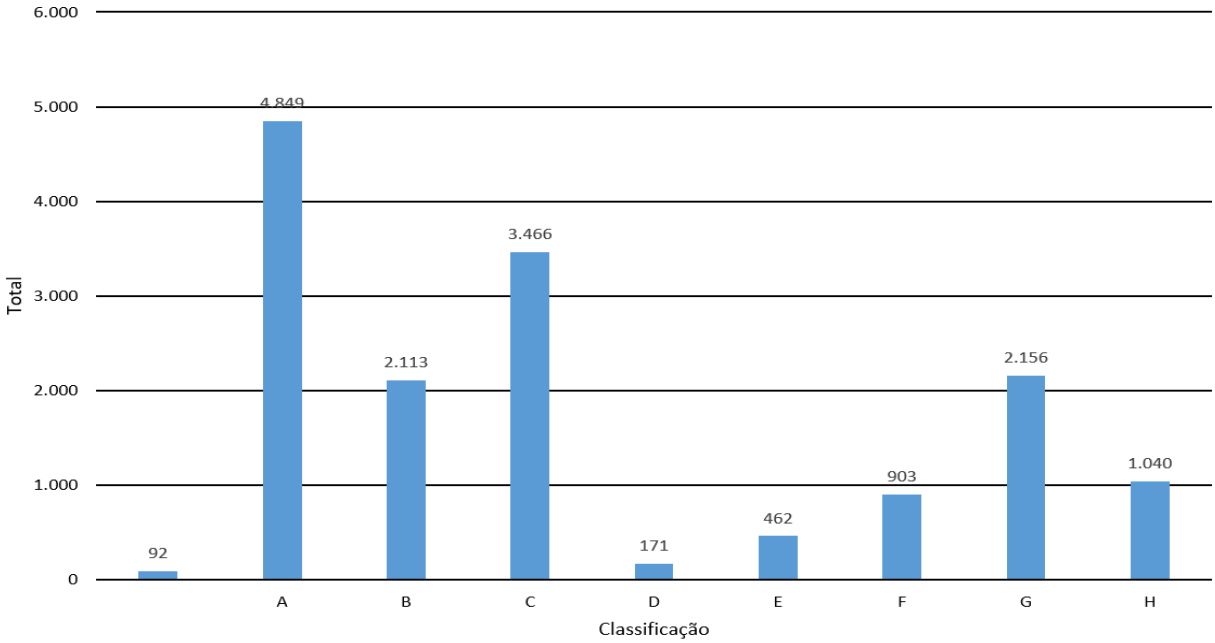


Fonte: Elaboração do autor.

As patentes informadas nos currículos cadastrados na Plataforma Lattes, foram depositadas entre os anos de 1975 a 2018, tendo maior concentração entre os anos de 2002 e 2016 destacando um crescimento no número de depósitos de patentes entre os anos 2000 e 2018. Ressalta-se que o conjunto de currículos utilizados na análise foram coletados em 2019 o que justifica a ausência de patentes nos últimos anos é possível queda brusca no ano de 2018, já que algumas patentes podem não terem sido cadastradas por seus proponentes.

Diante disso, várias análises podem ser viabilizadas para melhor compreensão do conjunto de dados recuperado. Cada patente de acordo com sua natureza e finalidade recebe uma classificação de acordo com o sistema internacional de classificação de patentes o IPC (*International Patent Classification*), a maior parte das patentes produzida pelos pesquisadores com informação extraída dos currículos, cerca de 32%, são classificadas como "A-Necessidades humanas". A Figura 4 apresenta um gráfico com as classificações das patentes informadas nos currículos da Plataforma Lattes.

Figura 4 – Classificação das patentes registradas nos currículos da Plataforma Lattes



A - Necessidades humanas; B - Operações de processamento, transportes; C - Química; Metalurgia; D – Têxteis, Papel; E – Construções fixas; F – Engenharia mecânica, Iluminação, Aquecimento, Armas, Explosões; G – Física; H – Eletricidade.

Fonte: Elaboração do autor.

Em um contexto geral, as patentes brasileiras em sua maioria recebem as classificações, A - Necessidades humanas (32%); B - Operações de processamento, transportes (14%); e C – Química (23%).

Diante do exposto, diversas novas análises poderão ser realizadas, principalmente pela integração dos repositórios de dados da Espacenet e dos dados registrados nos currículos da Plataforma Lattes. Tais análises serão importantes para melhor compreender como tem evoluído a produção técnica no país e qual o perfil dos pesquisadores que têm depositado patentes.

4 CONSIDERAÇÕES FINAIS

A partir dos resultados obtidos pelo conjunto de dados extraídos foi possível verificar a grande viabilidade e valor científico em adotar informações de patentes como fonte de dados para análises acerca da produção técnica de um país, região ou área do conhecimento, caracterizando como de suma importância para compreender o cenário tecnológico nacional. O grupo de patentes brasileiras identificada na Espacenet se caracteriza como uma parcela significativa de todo o conjunto de dados cadastrados no INPI, tendo em vista a alta

complexidade em identificá-las na Espacenet, devida a falta de um padrão de conversão dos números de depósito de patente registradas até o ano de 2011.

Avaliar a representatividade das patentes cadastradas nos currículos da Plataforma Lattes em repositórios internacionais se apresenta como uma alternativa importante, tendo em vista, que tal identificação poderá inclusive viabilizar uma validação dos dados registrados nos currículos. Apenas 72.256 registros de patentes foram identificados no conjunto composto por mais de 6.8 milhões de currículos analisados. Poucos currículos apresentam informações de patente, ou seja, apenas 1% dos currículos tem registros de patentes, destacando a necessidade de um estudo para melhor compreender este tipo de produção.

REFERÊNCIAS

AMADEI, J. R. P.; TORKOMIAN, A. L. V. As patentes nas universidades: análise dos depósitos das universidades públicas paulistas. **Ciência da Informação**, v. 38, n. 02, p. 9–18, 2009.

CATIVELLI, A. S. **Indicadores métricos de valor de patentes: construção de um Índice de Valor utilizando as patentes verdes brasileiras**. Tese (Doutorado em Ciência da Informação) — Universidade Federal De Santa Catarina - Centro De Ciências Da Educação,2020.

DIAS, T. M. R. **Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes**. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, setembro 2016.

ESPECENET. **Espacenet patent search**. 2021. Disponível em: https://worldwide. espacenet.com/patent/.

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.

UECE, U. F. do C. **INPI - Saiba mais sobre a nova numeração nos pedidos da DIRPA e da DICIG**. 2011. Acessado em 11 de maio de 2021. Disponível em: 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.

VALMIRA, S. P. M. M.; PERUCCHI. **Universidades e a produção de patentes: tópicos de interesse para o estudioso da informação tecnológica**. Perspectivas em Ciência da Informação, v. 19, n. 2, p. 15–36, apr 2014.