Patentes como fonte de dados para análise sobre a produção técnica

Patents as a source of data for analysis on technical production

Raulivan Rodrigo da Silva[1], Thiago Magela Rodrigues Dias[2]

(1) CEFET/MG, R. Álvares de Azevedo, 400 - Bela Vista, Divinópolis - MG, 35503-822, raulivan@cefetmg.br.

(2) CEFET/MG, R. Álvares de Azevedo, 400 - Bela Vista, Divinópolis - MG, 35503-822, thiagomagela@cefetmg.br.

Resumo:

Este trabalho busca contribuir com a compreensão do cenário tecnológico nacional, utilizando dados provenientes de patentes como objeto de análise. Objetivando contribuir com o projeto BRCRIS do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), propõem-se uma estratégia para coleta de patentes depositadas no Instituto Nacional da Propriedade Industrial (INPI), além de identificar proponentes de patentes na base curricular da Plataforma Lattes. A metodologia proposta, consiste em coletar as patentes depositadas no Brasil utilizando um repositório internacional de publicação de dados de patentes, a Espacenet, que contém dados de patentes de mais de 70 países. Posteriormente, por meio do conjunto ferramental proposto, obtêm-se os currículos da Plataforma Lattes que possuem dados de patentes informada, viabilizando certificar com dados da Espacenet um conjunto com 31.816 registros com informações de patentes em 16.445 currículos da Plataforma Lattes.

Palavras-chave: Patente; BRCRIS; Espacenet; Plataforma Lattes; Coleta de dados.

Abstract:

This work seeks to contribute to the understanding of the national technological scenario, using data from patents as an object of analysis. Aiming to contribute to the BRCRIS project of the Brazilian Institute of Information in Science and Technology (IBICT), a strategy for collecting patents deposited at the National Institute of Industrial Property (INPI) is proposed, in addition to identifying patent applicants in the curricular base of the Platform Lattes. The proposed methodology consists of collecting patents deposited in Brazil using an international repository for publishing patent data, Espacenet, which contains patent data from more than 70 countries. Subsequently, through the proposed toolkit, the Lattes Platform curricula that have informed patent data are obtained, making it possible to certify with Espacenet data a set with 31,816 records with patent information in 16,445 Lattes Platform curricula.

Keywords: Patent; BRCRIS; Espacenet; Lattes Platform; Data collect.

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. Com o advento das redes de compartilhamento de informações por meio da internet, é possível deparar com um expressivo volume de dados advindos produções científicas.

Mediante ao exposto, é importante mensurar toda essas informações produzidas para acompanhar o progresso científico e tecnológico, bem como contribuir para sua evolução. Os Estudos Métricos da Informação são uma das áreas de interesse da Ciência da Informação e tem como foco a identificação e avaliação da informação, seu alcance, influência e impacto, de acordo com Cabrini e Gracio (2011), os “Estudos Métricos” são constituídos por um conjunto de estudos relacionados à avaliação da informação produzida.

De acordo com o foco de interesse, da natureza da informação e do objeto de análise, os ramos dos estudos métricos podem ser classificados como Bibliométricos, Informétricos ou Infométricos, Cientométricos, Cibermétricos, Webométricos, Patentométricos e Arquivométricos (CURTY; DELBIANCO, 2020).

No contexto da produção técnica, documentos de patentes se apresentam como uma rica fonte de informação tecnológica. A compreensão do estado da técnica da arte por meio de documentos de patentes, consequentemente apresenta um cenário mais assertivo a respeito de tendências tecnológicas, setores promissores, bem como, a possibilidade de

novas tecnologias (NASCIMENTO; SPEZIALI, 2020). Os estudos e análise de documentos de patentes permitem identificar o conhecimento científico e convertê-lo em conhecimento tecnológico.

Conforme apontam Nascimento e Speziali (2020) o mapeamento de tecnologias utilizando informações contidas em documentos de patentes, é pouco explorado no Brasil, salientando que, por fazer parte de áreas estratégicas de muitas empresas, relatórios de mapeamentos tecnológicos mais completos não são disponibilizados de forma gratuita para consulta. Calzolaio et al. (2018) e Tanaka e Inui (2016) concordam que os dados de patentes contêm informações valiosas para análises técnicas, entretanto, esta é uma área pouco explorada também pelas universidades. A prospecção tecnológica é uma área de estudo recente principalmente no Brasil, que possui uma literatura limitada sobre o tema. Serão aceitos estudos concluídos e em andamento, desde que tenha resultados e conclusões, mesmo que parciais.

Nessa conjuntura, o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), desde 2014 está desenvolvendo o *Brazilian Current Research Information System* (BrCRIS), um ecossistema de sistemas, com objetivo de reunir dados de diversas fontes tais como dados de Projetos de pesquisa, financiamento, pesquisadores, infraestrutura de pesquisa, instituições de pesquisa e seus outputs em C&T, constituindo um conjunto de dados para criação de sistemas de recomendações em dados abertos.

2 Objetivos

O objetivo principal deste trabalho é apresentar uma estratégia de coleta de patentes depositadas no Instituto Nacional da Propriedade Industrial[1] (INPI) no repositório internacional Espacenet[2]. Ensejando contribuir com o desenvolvimento do BRCRIS, certificando os dados de patentes, cujo seus respectivos proponentes, informaram registros de patentes em seus currículos da Plataforma Lattes.

3 Procedimentos Metodológicos

A metodologia foi dividida em três partes distintas, a saber, (1) Coleta de dados de Patentes; (2) Coleta dos dados da Plataforma Lattes; e por fim (3) Tratamento e Seleção dos dados.

A primeira parte consiste na coleta de patentes brasileiras disponíveis na Espacenet, ou seja, patentes que foram depositadas no INPI até o ano de 2021 e disponibilizadas para consulta na Espacenet. A coleta dos dados foi realizada no primeiro semestre do ano de 2022. Para otimizar o processo de coleta foi desenvolvido um script utilizando a linguagem de programação Python, para acessar a OPS (*Open Patent Services*), um serviço web que a Espacenet disponibiliza para fornecer acesso à base de dados de patentes. Ressaltando que para consumir tais serviços é necessário realizar um cadastro no site da Espacenet para obter as credenciais de acesso (Espacenet, 2020). Fazendo uso do serviço de busca, é possível pesquisar as patentes depositadas em um determinado período, por exemplo, para consultar patentes brasileiras publicadas no período de janeiro de 2021 à dezembro de 2021, se faz necessário construir a consulta da seguinte forma: *search?q=pn=BR and pd="20210101 20210131",* em que *search* é serviço de consulta de patente. O parâmetro *q* que recebe os critérios da consulta, já no parâmetro *pn* é informado o código do país no número da publicação e por fim, utilizando o conector "*and*" é informado o parâmetro *pd* em que é informado o período de publicação da patente, destacando a data inicial e final que devem ser informadas no formato AAAAMMDD (ano, mês, dia).

Após obter a lista de patentes brasileiras disponíveis na Espacenet, se fez necessário realizar o obter dos dados das patentes, usando o serviço "*publication/epodoc/{número-de-depósito }/biblio*", onde foi informado o número de depósito de cada patente no formato EPODOC, este último é uma regra de formatação aplicado no número de identificação da patente.

[1] <https://www.gov.br/inpi/pt-br>

[2] <https://worldwide.espacenet.com>

Todas as patentes foram armazenadas em arquivos no formato .json (*JavaScript Object Notation*), em um diretório denominado "PATENTESBR". O nome de cada arquivos é formado pelo número da patente o qual armazena os dados, por exemplo “BR0107786A.json”, com base nesta informação, foram criados sub diretórios para armazenar as patentes, cujo o nome do subdiretório é composto pelos 4 primeiros caracteres do nome do arquivo, de acordo com o exemplo citado, o nome do diretório a qual ele pertence é “BR01”.

Dando sequência, a segunda etapa foi coletar currículos registrados na Plataforma Lattes que possuem informações de patentes, como o número do pedido de depósito ou o título da patente. O processo de coleta e seleção dos dados curriculares da Plataforma Lattes foi realizado por meio do framework *LattesDataXplorer* (DIAS, 2016). O framework possui um conjunto de técnicas e métodos responsáveis por coletar, selecionar, tratar e analisar os dados da Plataforma Lattes.

O extrator coleta os currículos e os armazena no formato XML em pastas, identificadas de 00 a 99. O nome da pasta selecionada para armazenar o arquivo e o nome do mesmo são definidas de acordo com seu número único de identificador de 16 dígitos, os dois primeiros números do identificador correspondem ao nome da pasta e os 14 dígitos restantes correspondem ao nome do arquivo salvo. A coleta dos currículos foi realizada no primeiro semestre de 2022.

Nos currículos da Plataforma Lattes o número de depósito de patente, bem como as demais informações, são inseridas pelo o próprio pesquisador, o que geralmente ocasiona uma falta de padrão no registro das informações. Portanto, se faz necessário implementar o terceiro passo que consiste em tratar os números de depósito informado nos currículos coletados para possibilitar a identificação dos mesmos na base coletada na Espacenet. Este procedimento consiste em realizar uma sequência de passos, a saber: o “passo-1” da consiste na remoção da formatação do número de depósito, removendo todos os caracteres especiais como ponto, vírgula, símbolos, dentre outros. Em continuidade no tratamento do número de depósito, no “passo-2” é realizada a remoção do último dígito que compõem o número. Feito isto, o “passo-3” consiste em uma busca aproximada pelo o número de depósito já tratado, esta busca aproximada consiste em pesquisar nos arquivos de patentes o número de depósito usando como critério de seleção: %[número tratado]%, em que as porcentagens representam qualquer caractere, ou seja, qualquer que seja o valor antes e/ou depois do número tratado será considerado como resultado satisfatório da busca. Caso tenha localizado alguma patente que atenda aos critérios de busca no “passo-3”, é realizado o “passo-4 ”, em que é verificado se o nome do pesquisador consta na lista de inventores da patente, para isso, usa-se o nome do pesquisador conforme foi informado em seu currículo, a exemplo “Raulivan Rodrigo da Silva”, posteriormente o nome é dividido por espaços formando uma lista, com base no exemplo dado, a lista ficaria com 4 elementos “[Raulivan, Rodrigo, da, Silva]”, feito isto, verifica se todos os elementos da lista correspondem aos nome dos inventores, respeitando a mesma grafia e ordem de ocorrência, ignorando outros possíveis termos existentes no nome dos inventores. Caso tenha encontrado, finaliza-se o processo considerando o número informado pelo pesquisado como válido, caso não tenha encontrado, finaliza-se o processo considerando o número informado pelo pesquisador como inválido. Ainda no “passo-3” existe um fluxo alternativo, em que, caso não tenha localizada uma patente é investigado se o número utilizado como critério de busca inicia com alguns dos prefixos “CI”,”DI”,”UM” ou “PI”, em caso negativo, finaliza o processo considerando o número informado inválido, mas em caso afirmativo, é feita a substituição do prefixo identificado por “BR”, voltando assim ao “passo-3” dando continuidade no fluxo já estabelecido.

Para automatizar o processo de validação, foi implementado um algoritmo utilizando a linguagem de programação Python.

4 Resultados

Os dados coletados da Espacenet totalizaram 858.622 registros de patentes, contendo 1.914.866 nomes de depositantes e 4.386.733 nomes de inventores, cerca de 25,6 Gb (Gigabyte) de dados. O quantitativo de patentes coletadas na Espacenet corresponde aproximadamente a 90,83% do conjunto de patentes depositadas no INPI.

Para apresentar um panorama da evolução tecnológica nacional, foi realizada a análise dos depósitos anuais de patentes, apresentando dados entre 1972 a 2021, data de registro da patente mais antiga até o último ano de análise. A **Figura 1 - Apêndice A** apresenta a evolução temporal no número de depósito de patentes, ressaltando que seis patentes não contém a informação de ano de depósito.

Cada patente de acordo com sua natureza e finalidade recebe uma classificação de acordo com o sistema internacional de classificação de patentes, com base nestas classificações é possível compreender quais áreas do conhecimento têm gerado o maior número de patentes. Para compreender melhor este cenário foi compilado no gráfico, apresentado pela **Figura 2 - Apêndice A**, as classificações recebidas pelas patentes brasileiras consideradas neste estudo.

Dentro do contexto da Plataforma Lattes, os resultados foram obtidos mediante a análise dos registros de patentes contidos nos currículos cadastrados. Atualmente a Plataforma Lattes é composta por mais de 7.4 milhões de currículos, que abrange indivíduos nos diversos níveis de formação acadêmica, no entanto, somente 29.516 possuem informações de patentes registradas, menos de 1% de toda a base de dados curriculares. Os 29.516 currículos possuem juntos um total 72.256 registros com informações de patentes, contudo, não foram todos considerados, apenas 31.816 registros foram devidamente identificados na base de dados coletada na Espacenet, totalizando 16.445 currículos. O restante não foi possível identificar na base de dados da Espacenet aplicando as estratégias definidas neste estudo. De encontro com objetivo deste estudo, todo conjunto de dados de patentes dos 16.445 currículos foram incorporados ao projeto BRCRIS no segundo semestre de 2022.

5 Conclusão ou Considerações Finais

Este estudo abordou a perspectiva da ciência da informação, expondo os dados provenientes de patentes como uma fonte confiável e ampla no que se refere ao desenvolvimento tecnológico nacional.

Neste contexto, mediante aos resultados obtidos, é possível concluir que todos os objetivos apresentados neste estudo foram alcançados. A estratégia de coleta de dados de patentes proposta, viabiliza manter a base de dados sempre atualizada, executando a mesma sempre atualizando o período de depósito desejado, permitindo obter os dados de patentes de anos subsequentes da mencionada neste estudo. Um fato relevante a ser considerado, que o processo de coleta não é um processo rápido de ser executado, a Espacenet impõe limites de coletas mensais, que quando são atingidos, é necessário parar a coleta e esperar a próxima semana, pois os limites são renovados a cada domingo. Outro ponto, que foi identificado durante a realização deste estudo, é que é recomendado executar a coleta fora do horário comercial, pois durante o horário comercial, quando se realiza muitas requisições à API da Espacenet, sua conta é bloqueada por algumas horas, voltando a funcionar corretamente.

Já considerando os dados da Plataforma Lattes, apenas cerca de 1% de todos os currículos da base de dados da Plataforma Lattes, possuem informações sobre o depósito de patentes, base composta por mais de 7.4 milhões de currículos. Do conjunto de registros de patentes recuperados dos currículos, nem todos puderam ser validados na Espacenet devido à inconsistência nos dados registrados, notabilizando a necessidade da existência de mecanismos de validação e certificação dos dados patentários informados pelos os proprietários dos currículos. Contudo, foi possível contribuir com o desenvolvimento do projeto BRCRIS, identificando proponentes de patentes na base de dados curriculares da Plataforma Lattes, consequentemente

fornecendo os dados das patentes identificadas.

Referências

CABRINI, E. F. T. de O. M. C.; GRACIO. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base scopus. **Perspectivas em Ciência da Informação**, v. 16, p. 16–28, out. 2011. Disponível em: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362011000400003&lng=pt&tlng=pt>. Acesso em: 16 set. de 2022.

CALZOLAIO, A. P. A. E. et al. Mapeamento dos registros de propriedade intelectual (patentes) na universidade federal do rio grande do sul. **Revista Brasileira de Gestão e Inovação**, v. 6, n. 1, p. 44–70, 2018. Disponível em: <http://www.ucs.br/etc/revistas/index.php/RBGI/article/view/5860>. Acesso em: 16 set. de 2022.

CURTY, N. R. R. G.; DELBIANCO. As diferentes métricas dos estudos métricos da informação: evolução epistemológica, inter-relações e representações. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 25, p. 01–21, 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, set. 2016.

NASCIMENTO, M. G. Raphael da S.; SPEZIALI. Patentometria: a utilização de dados
contidos em patentes como mecanismo de análise da predominância tecnológica dos nits.
IV Encontro Internacional de Gestão, Desenvolvimento e Inovação, nov. 2020.

NASCIMENTO, M. G. Raphael da S.; SPEZIALI. Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits. **IV Encontro Internacional de Gestão, Desenvolvimento e Inovação**, nov. 2020.

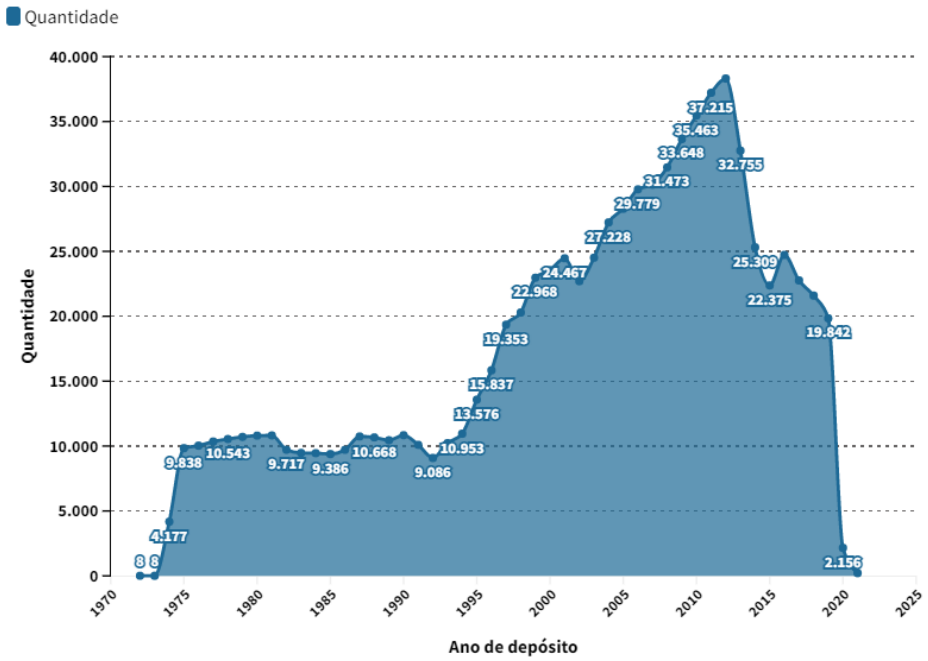
Espacenet. OPS. Open Patent Services RESTful Web Services. 1.3.16. ed. Disponível em: <https://www.epo.org/searching-for-patents/data/web-services/ops.html>. Acesso em: 13/01/2022.

PINTO, Adilson Luiz; SEGUNDO, Washington Luís Ribeiro de; QUONIAM, Luc; DIAS, Thiago Magela Rodrigues. Atas do V Congresso ISKO Espanha-Portugal: BRCRIS. In: **Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde**. 2021. cap. The brazilian current research information system, p. 319-330. ISBN 978-989-566-137-4. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=8411204>. Acesso em: 16 de set. 2022.

TANAKA, T. Y.; INUI. Preliminary study on why university researchers do not utilize patente information for their academic research in the field of science and engineering in japan. **Portland International Conference on Management of Engineering and Technology (PICMET)**, p. 1609–1618, 2016.

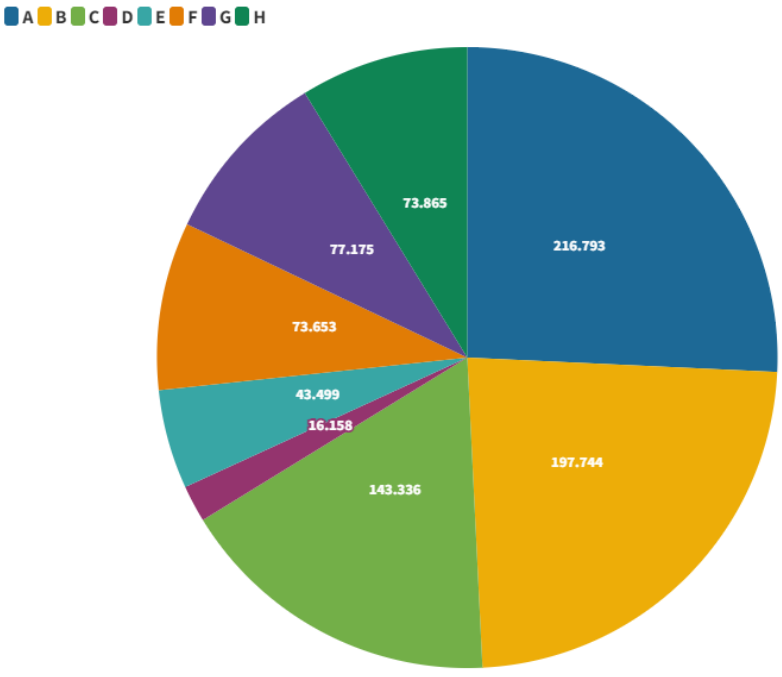
Apêndice A – Figuras

Figura 1 – Evolução temporal do depósito de patentes por ano



Dados da Pesquisa, 2022.

Figura 2 – Patentes por classificação



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.
Dados da Pesquisa, 2022