Uma estratégia para a identificação e extração de dados de patentes brasileiras

Raulivan Rodrigo da Silva
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Divinópolis, Brasil
raulivan@cefetmg.br

Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Divinópolis, Brasil
thiagomagela@cefetmg.br

Washington Luís Ribeiro de Carvalho Segundo
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Distrito Federal, Brasil
washingtonsegundo@ibict.br

***Abstract* — The main objective of this article is to present a strategy for the identification and extraction of data from Brazilian patents, such as title, abstract, filing date, publication date, inventors, owners, among others. Thus, enabling the construction of a local database of Brazilian technical production, enabling analysis of the large volume of data in a shorter time, since the analysis will be local and not in online patent repositories. As well as the solution for several limits imposed by online repositories, among them we can mention the limit on the volume of data access and internet connectivity. Using as main sources the National Institute of Industrial Property (INPI) and the international patent repository Espacenet, of recognized international relevance.**

***Resumo* — Este artigo tem como principal objetivo apresentar uma estratégia para a identificação e extração de dados de patentes brasileiras, tais como título, resumo, data de depósito, data de publicação, inventores, proprietários dentre outras. Viabilizando assim a construção de uma base de dados local da produção técnica brasileira, possibilitando análises do grande volume de dados em um menor tempo, uma vez que a análise será local e não em repositórios online de patentes. Bem como a solução para diversos limites impostos por repositórios online, dentre eles podemos citar o limite no volume de acesso à dados e conectividade com a internet. Utilizando como principais fontes o Instituto Nacional da Propriedade Industrial (INPI) e o repositório internacional de patentes Espacenet, de reconhecida relevância internacional.**

***Palavras-chave*—Coleta; Patente; INP; Espacenet.**

I. INTRODUÇÃO

Atualmente diversos repositórios disponíveis na internet possibilitam a pesquisa de produções científicas publicadas, a saber, DBLP (*Digital Bibliography Library Project*), ArnetMiner, Google Scholar, Microsoft Academic Search e a Plataforma Lattes do CNPq, sendo essa última um instrumento extremamente rico em dados para estudos sobre a produção científica e técnica brasileira. 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 internacional. A Espacenet, que viabiliza consultar em um único repositório patentes de aproximadamente 70 países, incluindo o Brasil, se destaca tendo em vista a quantidade de dados disponibilizada [2].

Embora existam diversos repositórios de consulta de patentes, estes apresentam limitações durante a consulta dos dados, como a consulta de um grande volume de dados, limite de tráfego de dados, limitam acesso via programação de boots, permitindo somente o acesso humano o que torna a análise onerosa[1].

Diante disso, este trabalho visa apresentar uma alternativa para viabilizar a análise de patentes. Uma estratégia para a identificação e extração de dados de patentes brasileiras, possibilitando a construção de uma base de dados local de patentes, onde permitirá ao pesquisador elaborar suas análises com maior liberdade em manipular, tratar e formatar os dados conforme a necessidade.

II. 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 [4].

Foi realizada a coleta de informações referente a documentos de patentes depositadas no INPI no período de 01/01/1900 à 31/12/2020. Assim, 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 que irá compor a base de dados, buscando garantir a consistência dos dados coletados.

Nas próximas seções, são apresentados: o processo de aquisição (A); e o processo de organização e tratamento dos dados (B).

A. 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.

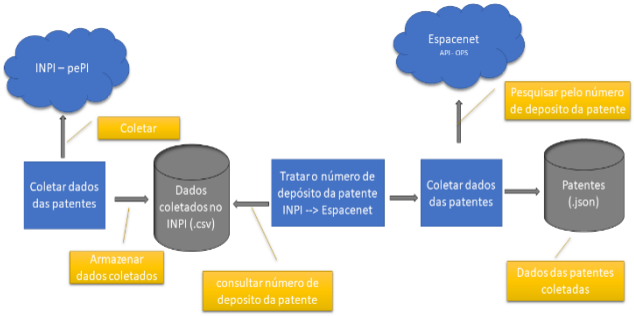

Fig. 1. Visão geral da coleta de dados

1) Coleta de dados no INPI: 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. Após acessar a ferramenta pelo método identificação anônima, foi selecionada a opção de "Patentes", onde é apresentada uma página contendo opções de pesquisa, foi selecionada a opção "pesquisa avançada" para exibir mais critérios de pesquisa.

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. De forma ilustrativa, se executarmos a coleta manual de todas as patentes, isso demandaria um esforço humano muito grande, considerando que levasse aproximadamente cerca de 10 minutos para acessar os detalhes cada patente e armazenar as informações de interesse em um processador de planilhas eletrônicas, levaria cerca de 143.434 horas, dedicando 8 horas por dia, levaria cerca de 17.942 dias. 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;
    b. 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".
    c. Armazenar as informações recuperadas em um arquivo CSV (Comma-separated-values);
    d. Voltar a listagem de patentes;
5. 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. Zhao (2017) define *web scraping*, que em tradução para português, raspagem web, como uma técnica de extração de dado em páginas disponíveis na WWW (*World Wide Web*) e armazená-lo em arquivo ou em banco de dados para posterior análise, podendo ser realizado manualmente por um usuário ou automaticamente por um robô (*web crawler*) [6]. *Web crawler*, é um algoritmo usado para encontrar, ler e indexar páginas de um site. *Web scraping* engloba um grande conjunto de técnicas de programação e diversas tecnologias, tais como, a análise de dados, *parsing* de idiomas naturais e segurança da informação, entre outras [3].

Durante os testes do algoritmo desenvolvido foi possível identificar uma limitação na abordagem estabelecida, 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", conseguindo assim, em apenas 0,5% do tempo, se comparado ao processo manual, para coletar as informações das patentes. 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.

2) Coleta de dados na Espacenet: 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 [2].

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. Esta 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, o primeiro utilizado para patentes mais antigas e atualmente é adotado um segundo 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 [5].

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. A figura 2 apresenta graficamente a composição do formato.

Fig. 2. Formato antigo de numeração de patentes.

O novo formato estabelecido visa atender à política de integração internacional do INPI atendendo os padrões sugeridos internacionalmente pela Organização Mundial de Propriedade Intelectual (OMPI) . 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 (tabela I), **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 (tabela II). A figura 3 apresenta graficamente a composição do formato atual de numeração de pedidos de depósito de patentes.

TABELA I
NATUREZA DA PROTEÇÃO

Código	Tipo de patente	Descrição
10	Patentes de Invenção	Pedido de patente de invenção depositado no INPI e via CUP (antigo PI)
11	Patentes de Invenção	Pedido de patente de invenção depositado via PCT
12	Patentes de Invenção	Pedidos divididos (antigo PI)
13	Patentes de Invenção	Certificado de adição (antigo C1, C2, etc)
14-19	Patentes de Invenção	Para atender necessidades da DIRPA
20	Modelo de Utilidade	Pedidos depositados por nacionais e via CUP (antigo MU)
21	Modelo de Utilidade	Pedidos depositados via PCT (antigo MU PCT)
22	Modelo de Utilidade	Pedidos divididos (antigo MU)

23-29	Modelo de Utilidade	Para atender necessidades da DIRPA

TABELA II
EXEMPLOS DE CÓDIGOS DE PUBLICAÇÃO

Código	Descrição
A1	Publicação do pedido de patente
A2	Publicação do pedido sem o relatório de busca
A3	Publicação do pedido com o relatório de busca
B1	Publicação da patente concedida
B2	Republicação da patente, por estar ilegível

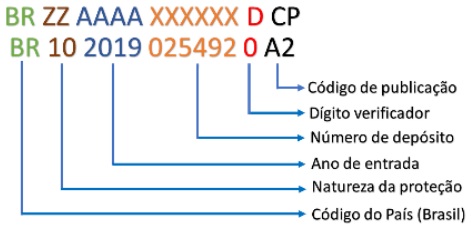

Fig. 3. Novo formato de numeração

Após uma série de experimentos, baseados em tentativa e erro, removendo e incluindo caracteres no número de pedido de depósito de patente foi possível definir uma sequência de passos para tratar os números de pedido de depósito coletados no INPI, para identificar a patente correspondente na Espacenet. As regras definidas foram as seguintes:

1. Regra R001: Caso o número inicie com "BR", será removido os espaços entre os caracteres e desconsiderados os 3 últimos caracteres. Exemplo: BR 10 2019 025492 0 A2 → BR102019025492.
2. Regra R002: Caso o número não inicie com "BR", mas inicie com alguns dos seguintes prefixos: "C1", "DI", "MI", "MU" ou "PI", será adotada uma estratégia de força bruta:
    a. Regra R002-01: A primeira tentativa é incluir o prefixo " BR ", desconsiderar todos os caracteres após o separador " - " e remover todos os espaços entre os caracteres. Exemplo: PI 0000061-2 B1 → BRPI0000061. Caso não apresente sucesso, iniciar o passo descrito na letra "b"
    b. Regra R002-02: A segunda tentativa consiste em substituir os dois primeiros caracteres por " BR ", desconsiderar todos os caracteres após o separador " - " e remover todos os espaços entre os caracteres. Exemplo: PI0300001-0 A2 → BR0300001. Caso não apresente sucesso, iniciar o passo descrito na letra "c".
    c. Regra R002-03: A terceira tentativa consiste em executar os mesmos passos definidos na regra R002-02 , incluindo o caractere que estiver dentro do intervalo de " A-Z " , que houver após o separador. Exemplo:PI0300001-0 A2 → BR0300001A. Caso não apresente sucesso, iniciar o passo descrito na letra "d".
    d. Regra R002-04: A quarta tentativa, consiste em executar os mesmos passos definidos na regra R002-03, incluindo todos os caracteres que existirem após o caractere dentro do intervalo de "A-Z " , que existir após o separador. Exemplo: PI0300001-0 A2 → BR0300001A2.

Neste processo é considerando sucesso somente quando é localizado uma única patente na Espacenet, não localizar a patente ou localizar mais de uma patente, será considerado como falha. Tomando como base o 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, com a seguinte estrutura de campos:

- INPI Number: Número do pedido de depósito de patente da forma em que foi coletado no INPI;
- Original number: Número do pedido de depósito de patente da forma em que foi coletado no INPI, porém sem todos os espaços;
- Rule 001: Resultado do processamento da regra R001;
- Rule 002-01: Resultado do processamento da regra R002-01;
- Rule 002-02: Resultado do processamento da regra R002-02;
- Rule 002-03: Resultado do processamento da regra R002-03;
- Rule 002-04: Resultado do processamento da regra R002-04;
- Json file name: Nome que será atribuído aos dados coletados das patentes quando os mesmos forem localizados na Espacenet.

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.

B. Organização e tratamento dos dados

Essa subseção tem como objetivo apresentar a organização dos documentos de patentes coletados na Espacenet, destacar e conceituar os principais elementos que o compõem.

Os documentos de patentes foram armazenados em documentos de extensão JSON, para facilitar a manipulação dos mesmos. Os arquivos foram armazenados em um diretório de nome "patentes-br", onde dentro desse diretório tem subpastas cujo os nomes contém os 4 primeiros dígitos de um número de patente, por exemplo: "BR10", "BR01", "BR20". A fim de organizar e agrupar as patentes para otimizar processos de consultas. A figura 4 ilustra a organização dos documentos de patentes.

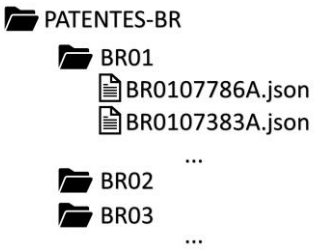

Fig. 4. Organização de pastas e arquivos de patentes

O arquivo JSON de patente possui diversas informações acerca da patente, dentre elas temos:

- Dados bibliográficos (metadados): Os dados bibliográficos contêm todos os documentos utilizados para pleitear a concessão da patente, como o relatório descritivo, o quadro reivindicatório, o resumo, desenho, dentre outros.
- Requerentes: Um Requerente ou Cessionário é a pessoa, ou organização, como empresas, universidades que formalizou um pedido de patente. O pedido de patente pode conter um ou mais requerentes, inclusive o requerente pode ser o próprio inventor.
- Inventores: O inventor representa uma ou mais pessoas as quais foram responsáveis pelo esforço intelectual associado à invenção. Diferente do nome do requerente, os nomes dos inventores raramente sofrem alterações ao longo do ciclo de vida de um pedido de patente, exceto para correções.
- Datas: As datas correspondem à momento de eventos significativos no ciclo de vida de um pedido de patente. As três principais datas são a data de prioridade, a data de depósito e as datas de publicação.
- Classificações: A classificação de patente agrupa as patentes de acordo com a sua especificação técnica, facilitando a organização e recuperação de documentos de patentes pelos escritórios de propriedade intelectual e demais usuários.
- Citações: As citações consistem em relacionar informações contidas em diferentes documentos de patentes, estabelecer uma relação interna entre o documento citado com o citante.

III. ANÁLISE E DISCUSSÃO DOS RESULTADOS

Como resultado, 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, que poderá ser identificado e tratado em trabalhos futuros.

Com o sucesso da coleta das patentes também foi possível coletar, por meio da Espacenet, as famílias das patentes brasileiras, até o presente momento 21% das famílias foram coletadas e armazenadas em arquivos de extensão JSON.

Todos os dados coletados somam um total de 23,7 GB de dados sobre a produção técnica brasileira.

IV. CONCLUSÃO

A partir dos resultados obtidos foi possível validar a estratégia proposta por este trabalho, onde por meio do tratamento do número de pedido de depósitos de patentes, convertendo-os para o padrão internacional obteve sucesso ao identificar 83% das patentes brasileiras registradas na Espacenet. Bem como apontar grande viabilidade em trabalhar com dados locais, pela flexibilidade para manipular e tratar os dados, além do enorme 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.

Entretanto, trabalhos futuros podem ser desenvolvidos a fim de coletar 100% das patentes brasileiras registradas na Espacenet.

AGRADECIMENTOS

Pelo auxílio os autores agradecem ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET/MG).

REFERÊNCIAS

[1] 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. Disponível em: <https://lume.ufrgs.br/handle/10183/179853>. Acesso em: 24 mar. 2021

[2] ESPECENET. *Espacenet patent search*. 2021. Disponível em: <https://worldwide.espacenet.com/patent/>.

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

[4] 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.

[5] 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. Disponívelem:<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>.

[6] ZHAO, B. Web scraping.Springer International Publishing, p. 1–3, may 2017. Disponível em: <https://www.researchgate.net/publication/317177787_Web_Scraping>. Acesso em: 07 mai. 2021.