



# FRAMEWORK DE AUTOMAÇÃO DE COLETA DE DADOS BIBLIOGRÁFICOS DE PATENTES EM REPOSITÓRIOS DE ACESSO ABERTO

**Raulivan Rodrigo da Silva**

Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brasil | [raulivan@cefetmg.br](mailto:raulivan@cefetmg.br)

 <https://orcid.org/0000-0002-2740-1045>

**Thiago Magela Rodrigues Dias**

Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brasil | [thiagomagela@cefetmg.br](mailto:thiagomagela@cefetmg.br)

 <https://orcid.org/0000-0002-3090-9413>

**Washington Luís Ribeiro de Carvalho Segundo**

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasil | [washingtonsegundo@ibict.br](mailto:washingtonsegundo@ibict.br)

 <https://orcid.org/0000-0003-3635-9384>

**DOI:** 10.22477/xiv.biredial.401

**EJE TEMÁTICO:** Infraestrutura tecnológica

## RESUMEN

Este estudo tem como principal objetivo propor uma estratégia automatizada para a coleta de grandes volumes de dados bibliográficos de patentes em repositórios de acesso aberto. Para tanto, é adotada uma abordagem metodológica quantitativa e exploratória, com ênfase no desenvolvimento de estratégias eficazes para a pesquisa e recuperação automatizada de informações no repositório da Espacenet. Os resultados obtidos incluem o desenvolvimento de um framework implementado em Python, capaz de automatizar integralmente o processo de coleta e recuperação de dados. Essa solução estabelece um processo sistemático para extração de informações bibliográficas de patentes, bem como amplia a eficiência do processo, facilitando a obtenção de conjuntos de dados relevantes para análise e pesquisa. Dessa forma, o estudo contribui para o aprimoramento de estratégias computacionais voltadas à recuperação e ao tratamento de informações no campo da propriedade industrial.

**Palabras-clave:** Patentes; Automação; Coleta; Espacenet.

## ABSTRACT

(ver resumo em inglês com autores)

**Keywords:** Patents; Automation; Collection; Espacenet.

## INTRODUÇÃO

O campo da Propriedade Intelectual desempenha um papel crucial no estímulo à inovação e ao desenvolvimento tecnológico em diversas áreas do conhecimento. Os documentos de patentes se destacam como fontes valiosas de informações sobre inovações tecnológicas, tendências de mercado e estratégias competitivas das organizações.

Nesse ambiente dinâmico e em constante mudança, acompanhar o desenvolvimento tecnológico de um setor ou de um grupo de empresas tornou-se um pré-requisito para aquelas que desejam se destacar no campo da inovação. Dentre as várias formas de monitorar esse desenvolvimento, segundo Nascimento e Speziali (2020), o uso das informações tecnológicas contidas em documentos de patentes é uma excelente alternativa. A compreensão do estado da arte por meio desses documentos apresenta um cenário mais preciso sobre a identificação de tendências tecnológicas e setores promissores (Nascimento & Speziali, 2020). Conforme apontado por Sanz-Casado (2006), estudos e análises de patentes permitem identificar o conhecimento científico e convertê-lo em conhecimento tecnológico.

Nesta perspectiva, é coerente classificar os documentos de patentes como uma fonte rica e consistente de informações tecnológicas, pois os documentos de patentes exigem uma descrição minuciosa e detalhada das invenções, incluindo desenhos, esquemas e reivindicações que delimitam o escopo da tecnologia. Esses documentos destacam-se como recursos essenciais para pesquisadores e estudiosos da área de Propriedade Intelectual (Rezende *et al.*, 2023). Segundo Reymond e Quoniam (2018), os documentos de patentes contêm uma quantidade substancialmente maior de informações em comparação com aquelas encontradas em artigos científicos, ampliando assim sua relevância e utilidade.

Entretanto, a coleta de dados de patentes pode apresentar vários desafios devido à natureza específica e muitas vezes complexa dos documentos de patentes. Frequentemente, as patentes são redigidas em linguagem técnica e jurídica, o que pode dificultar a interpretação para aqueles que não estão familiarizados com o campo da Propriedade Intelectual. Além disso, o grande número de patentes registradas globalmente e a diversidade de áreas tecnológicas tornam a coleta de dados uma tarefa desafiadora. Muitos repositórios de patentes de acesso aberto disponíveis na internet oferecem recursos limitados para a recuperação de dados, o que muitas vezes exige o uso de repositórios pagos, tornando o processo de coleta mais demorado e oneroso. Além disso, com a constante concessão e atualização de novas patentes, manter os dados atualizados requer um esforço contínuo para garantir que as informações coletadas permaneçam precisas e relevantes.

Mesmo com o progresso tecnológico e a digitalização facilitando o acesso a documentos de patentes através de repositórios online, ainda persistem obstáculos consideráveis que impedem a utilização total desses dados pelos pesquisadores. Embora repositórios de acesso



livre, como o Espacenet ou Patentscope, sejam frequentemente usados, eles apresentam restrições técnicas na consulta e recuperação de patentes (Brandão, 2016).

Tendo em vista os fatos apresentados, é evidente que a coleta de documentos de patentes constitui uma tarefa complexa, com lacunas ainda a serem exploradas. Neste sentido, o presente estudo tem como objetivo explorar estratégias que viabilizem coletas mais abrangentes de documentos de patentes em repositórios de acesso aberto.

## FUNDAMENTAÇÃO TEÓRICA

Por definição, os documentos de patentes são de acesso público. Esta característica resulta de um dos princípios básicos do sistema de patentes: a divulgação da informação técnica em troca da concessão de um direito exclusivo de exploração comercial. Ao tornar a invenção acessível ao público, o detentor da patente contribui para o avanço tecnológico e científico, possibilitando que outros inventores e pesquisadores se apoiem no conhecimento compartilhado. A Lei no 9.279/96 (1996) (Lei de Propriedade Industrial) do Brasil, em seu Artigo 30, estabelece que os documentos de patente devem ser divulgados após o período de sigilo, assegurando que as informações sejam disponibilizadas ao público. Há diversos repositórios de patentes disponíveis na internet que permitem o acesso a documentos de patentes. Esses repositórios podem ser de acesso aberto ou comerciais, nacionais ou internacionais, entre outras classificações.

Os repositórios nacionais, também conhecidos como repositórios locais, estão vinculados aos escritórios de propriedade industrial de cada país, sendo comuns em nações que adotam a proteção da propriedade intelectual. Geralmente, esses repositórios disponibilizam apenas os documentos de patentes depositados em seus respectivos escritórios. Dentre eles, destacam-se sete repositórios nacionais, a saber:

- *European Patent Office (EPO)*;
- *Intellectual Property Índia (IPI)*, disponibiliza as patentes depositadas na Índia;
- *Japan Patent Office (JPO)* para as buscas das patentes depositadas no território japonês. Mantido pelo Escritório japonês de Patentes, com a possibilidade de se obter cópia dos documentos originais japoneses;
- *Korean Intellectual Property Office (KIPO)*, repositórios de patentes da Coreia do Sul;
- Pesquisa em Propriedade Industrial (pePI), é o repositório de documentos de patentes do escritório de patentes do Brasil vinculado ao governo federal, com acesso aberto e mantido pelo Instituto Nacional da Propriedade Industrial (INPI);
- *State Intellectual Property Office of China (SIPO)* para buscas em patentes depositadas na República Popular da China;

- e o *United States Patent and Trademark Office* (USPTO) o repositório de patentes depositadas nos Estados Unidos da América (EUA).

Em um contexto internacional, que abrange repositórios que disponibilizam acesso a documentos de patentes de diversos escritórios de propriedade industrial, é possível destacar, em concordância com Brandão (2016) e Pires *et al.* (2020), as seguintes plataformas: Espacenet, Latipat, Patentscope, Google Patents, Lens, Derwent Innovation Index e Orbit Intelligence.

A Espacenet é um repositório gerenciado pelo Escritório Europeu de Patentes (EPO), que oferece acesso aberto a mais de 149 milhões de documentos de patentes provenientes de diversos escritórios de Propriedade Intelectual de mais de 100 países, abrangendo o período de 1782 até a atualidade (Espacenet, 2025). Já o repositório Latipat é uma interface de pesquisa de acesso aberto, desenvolvida com tecnologia da Espacenet, destinada à busca de informações técnicas em espanhol e português em documentos de patentes de 18 países da América Latina e da Espanha (Latipat, 2025).

Por sua vez, Patentscope é o repositório de busca de patentes fornecido pela Organização Mundial da Propriedade Intelectual (OMPI) e foi lançado em 2005. O repositório Patentscope permite que usuários pesquisem e acessem informações sobre patentes de diversos países. Atualmente o repositório permite pesquisar 115 milhões de documentos de patentes, incluindo 4,8 milhões de pedidos de patente internacionais publicados (PCT) a partir de 1782 (Abrantes *et al.*, 2022).

O repositório Google Patents é um serviço oferecido pelo Google que permite pesquisar e acessar informações relacionadas a patentes concedidas e pedidos de patentes em todo o mundo. Lançado pela Google em 2006, o Google Patents possui um arcabouço de mais de 120 milhões de documentos de patentes, disponibilizado dados a partir de 1790, bem como documentos técnicos e livros indexados no Google Acadêmico e no Google Livros (Google, 2025).

Em sequência, o repositório Lens é uma plataforma online que oferece acesso aberto a registros relacionados a patentes, literatura científica, dados de pesquisa e inovação. O repositório pode ser considerado um agregador de metadados, com acesso a trabalhos acadêmicos com indexação de citações e acesso a uma coleção abrangente de dados bibliográficos de patentes, também com indexação de citações (The Lens, 2025).

Ainda no contexto de repositórios de patentes de abrangência internacional, a Clarivate Analytics fornece um repositório comercial chamado *Derwent Innovation Index*. Com dados provenientes do *Derwent World Patents Index* (DWPI), o repositório *Derwent Innovation Index* disponibiliza dados de 52 escritórios de patentes, 35 milhões de famílias de patentes e 74 milhões de documentos de patentes, publicadas desde 1963 (Pires *et al.*, 2020).

Por fim, o repositório *Orbit Intelligence*, uma solução dedicada a especialistas em informações de patentes de propriedade da Questel Co. Permite pesquisar, monitorar, analisar, categorizar e disseminar informações de patentes (Questel, 2025).



Contudo, é necessário ressaltar que apesar de apresentarem recursos semelhantes, cada repositório de patentes tem sua própria interface e algumas funcionalidades exclusivas. Apesar da ampla variedade de repositórios de patentes disponíveis, os autores Singh *et al.* (2016) concordam e destacam que nenhum deles abrange a totalidade das patentes publicadas no mundo.

## METODOLOGIA

O presente estudo possui natureza de pesquisa aplicada, desenvolvida com orientação epistêmica hipotético dedutiva alinhada com o problema e objetivo anteriormente apresentados. O foco está em gerar conhecimento prático voltado para a resolução de problemas específicos no campo da propriedade industrial, particularmente recuperação de dados bibliográficos de patentes.

Adicionalmente, a pesquisa adota uma orientação metodológica exploratória e descritiva, pois busca explorar estratégias e ferramentas para coletar e quantificar os dados.

Quanto aos procedimentos, a pesquisa combina métodos de pesquisa documental e bibliográfica, pois baseia-se na análise de documentos de patentes disponíveis publicamente, além da revisão da literatura pertinente à repositórios de patentes.

## DESENVOLVIMENTO

Para o processo de automação de coleta de dados bibliográficos de patentes será adotado o repositório Espacenet, devido o repositório disponibilizar um serviço web de coleta, denominado *Open Patent Services* (OPS). OPS é um serviço da web que fornece acesso aos dados armazenados no banco de dados do *European Patent Office* (EPO) por meio de serviços web utilizando 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 possibilidade consequentemente, torna viável o desenvolvimento de aplicativos e robôs de extração automática para baixar grandes volumes de dados.

Desta forma, fazendo uso do serviço de pesquisa disponível na OPS, é possível informar filtros para a seleção dos registros que se deseja recuperar. Usando a sintaxe *Contextual Query Language* ou simplesmente CQL, de acordo com as especificações definidas pela documentação (OPS, 2025). É importante destacar que, a Espacenet limita a recuperação a 500 registros por consulta. Isso significa que, caso a busca abranja mais de 500 registros, apenas os primeiros 500 serão retornados. Para contornar essa limitação e garantir a recuperação do máximo possível de registros, são necessárias consultas incrementais, organizadas de acordo com a

data de publicação das patentes. Por exemplo, para obter as patentes publicadas entre janeiro e dezembro de 2021, é preciso realizar uma consulta para cada dia do período.

Os requisitos técnicos para acessar os serviços da OPS estão descritos detalhadamente na documentação oficial fornecida pelo *European Patent Office*<sup>1</sup>. Com base nas informações presentes na documentação, Figura 1 ilustra o processo de consulta incremental de patentes.

<sup>1</sup> European Patent Office. Disponível em: <https://www.epo.org/en/searching-for-patents/data/web-services/ops>). Acesso em: 2 set. 2025.



Figura 1 - Pesquisa de Patentes na Espacenet

### Algoritmo 1: PESQUISA DE PATENTES NA ESPACENET

**Entrada:**  $pConsulta$ ,  $pDataInicial$ ,  $pDataFinal$

**Saída:** Conjunto de resultados da pesquisa de patentes

```

1  início
2  |  $vDataAtual \leftarrow pDataInicial$ ;
3  | while  $vDataAtual \leq pDataFinal$  do
4  |   |  $vRangeInicial \leftarrow 1$ ;
5  |   |  $vRangeFinal \leftarrow 100$ ;
6  |   |  $vContinuar \leftarrow \text{True}$ ;
7  |   | while  $vContinuar = \text{True}$  do
8  |   |   | /* Aguardar 5 segundos antes de continuar */
9  |   |   | Esperar(5);
10 |   |   |  $vResultado \leftarrow$ 
11 |   |   |   | pesquisar( $pConsulta$ ,  $vDataAtual$ ,  $vRangeInicial$ ,  $vRangeFinal$ );
12 |   |   |   | salvarResultado( $vResultado$ );
13 |   |   |   |  $vTotalResultados \leftarrow \text{retornaTotalResultados}(vResultado)$ ;
14 |   |   |   | if  $vTotalResultados > vRangeFinal$  then
15 |   |   |   |   |  $vRangeInicial \leftarrow vRangeInicial + 100$ ;
16 |   |   |   |   |  $vRangeFinal \leftarrow vRangeFinal + 100$ ;
17 |   |   |   | end
18 |   |   |   | else
19 |   |   |   |   |  $vContinuar \leftarrow \text{False}$ ;
20 |   |   |   | end
21 |   | end
22 |  $vDataAtual \leftarrow \text{incrementaDia}(vDataAtual)$ ;
23 end
24 fim

```

**Fonte:** Elaboração dos autores (2025).

O algoritmo apresentado é destinado à pesquisa de patentes no repositório Espacenet por meio dos serviços da OPS. Ele recebe três parâmetros como entrada:  $pConsulta$ , que repre-



sentam os critérios de buscas a serem considerados na pesquisa; *pDataInicial*, que é a data de inicial do período de publicação das patentes; e *pDataFinal*, que é a data final do período de publicação das patentes.

Na linha 2, o algoritmo começa inicializando a variável *vDataAtual* com a data inicial *pDataInicial*. Em seguida, entra no primeiro loop, que continua enquanto a data atual *vDataAtual* for menor ou igual à data final *pDataFinal*. Dentro do loop, são inicializadas as variáveis *vRangeInicial* e *vRangeFinal*, que definem o intervalo de paginação dos resultados. Inicializa também a variável *vContinuar* como verdadeira para controlar o segundo loop.

Na linha 7 o algoritmo entra no segundo loop que continua enquanto a variável *vContinuar* for verdadeira. O algoritmo espera por 5 segundos antes de continuar a execução, isso é útil para evitar que a OPS bloqueie as requisições. Na linha 9, o algoritmo realiza a pesquisa de patentes na Espacenet com base nos parâmetros fornecidos, incluindo a consulta *pConsulta*, a data atual *vDataAtual* e os intervalos de resultados *vRangeInicial* e *vRangeFinal*.

O resultado da pesquisa é armazenado na variável *vResultado*. O algoritmo chama a função *salvarResultado()* para salvar em arquivo o resultado da pesquisa. Na linha 12, verifica se o total de resultados *vTotalResultados* é maior que o valor final do intervalo de resultados *vRangeFinal*. Se for, atualizar os intervalos de pesquisa para a próxima página de resultados. Caso contrário, define a variável *vContinuar* como falsa para sair do segundo loop.

Por fim na linha 20, chama a função *incrementaDia()* para incrementar a data atual *vDataAtual* para o próximo bibliográficos das patentes.

o dia. Todo esse tratamento é devido ao fato de a OPS paginar os resultados. No retorno vem informado o total de registros que atenderam os critérios de seleção e apenas os 100 primeiros destes resultados, sendo necessários realizar novas consultas para buscar a próxima página de resultados até recuperar todos.

Contudo, o método *search* não retorna diretamente os dados bibliográficos, mas apenas os números de depósito das patentes que satisfazem os critérios de busca estabelecidos. Para obter os dados bibliográficos completos, é necessário utilizar um segundo método, o *published - data*, que permite a recuperação detalhada das informações bibliográficas de cada patente. Esse processo exige a abertura de cada arquivo de resultado, seguida da leitura e processamento de cada registro, a fim de recuperar os dados bibliográficos correspondentes a cada patente. A Figura 2 apresenta o pseudocódigo que descreve o fluxo do processo de coleta dos dados



**Figura 2** - Fluxo do processo de coleta dos dados bibliográficos das patentes

### Algoritmo 2: COLETA DE DADOS BIBLIOGRÁFICOS NA ESPACENET

**Entrada:** *pListaArquivosResultados*

**Saída:** Arquivo com dados bibliográficos das patentes

```

1 início
2   para cada vArquivo em pListaArquivosResultados faça
3     vResultados ← recuperarRegistros(vArquivo);
4     para cada vPatente em vResultados faça
5       vDadosBibliograficosPatente ←
6         recuperarDadosBibliograficos(vPatente);
7       salvarEmArquivo(vPatente);
8     fim
9 fim
  
```

Fonte: Elaboração dos autores (2025).

Conforme apresentado, o algoritmo recebe como entrada uma lista com o caminho de todos os arquivos de resultados, obtidos anteriormente no processo de pesquisa das patentes. O algoritmo utiliza uma estrutura de repetição para percorrer cada arquivo *vArquivo* na lista *pListaArquivosResultados*. Assim, para cada arquivo de resultado *vArquivo*, o algoritmo chama a função *recuperarRegistros()* para recuperar os registros de patentes contidos no arquivo. Esses registros são armazenados na variável *vResultados*.

Na linha 4, uma nova estrutura de repetição é declarada para percorrer cada patente *vPatente* nos resultados *vResultados*. Na linha 5, para cada patente *vPatente*, o algoritmo chama a função *recuperarDadosBibliograficos()* para recuperar os dados bibliográficos associados à patente, armazenando-os na variável *vDadosBibliograficosPatente*. E por fim, chama a função *salvarEmArquivo()* para salvar os dados da patente em um arquivo.

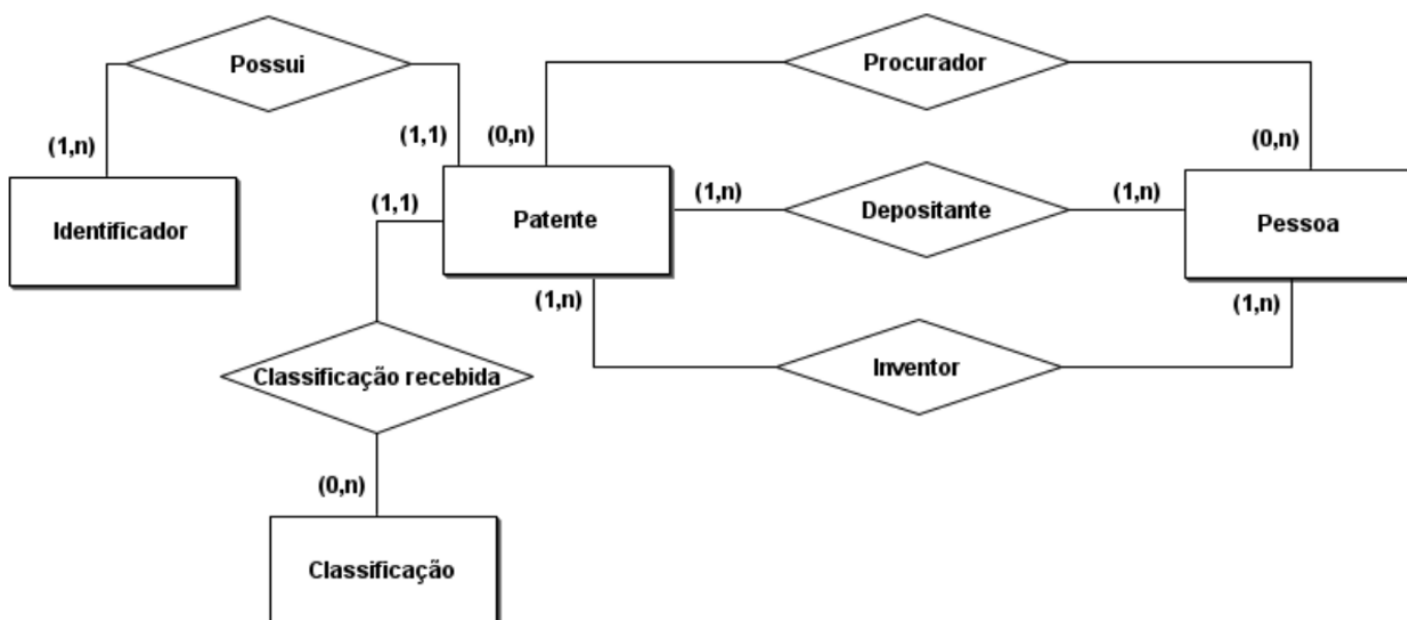
As patentes coletadas foram armazenadas em arquivos no formato .json, formato este que potencializa a manipulação dos dados. Entretanto, para extrair informações que dependem de relacionamento e/ou cruzamento de dados, exige maior esforço computacional, na construção de algoritmos que colem as informações necessárias dos arquivos. Como por exemplo, gerar um indicador de patentes por classificação, em que seria necessário percorrer todos os arquivos para coletar as informações de classificação e contabilizá-las. Com intuito de resolver este problema e viabilizar novas estratégias de processamento e cruzamento de

informações, foi também adotado o uso de um banco de dados relacional, baseado no Modelo Entidade Relacionamento introduzido por Peter Pin-Shan Chen, em 1976 (Date, 2003).

Para a construção do banco de dados local, foi adotado o SQLite, pois ele não exige a implementação de um servidor dedicado e dispensa configurações complexas. Essas características tornam o SQLite uma solução prática e de fácil manipulação, além de suportar até 280 terabytes de dados (The Sqlite Consortium, 2025).

Diante do exposto, fundamentado na estrutura dos arquivos coletados na Espacenet, foi estabelecido o esquema conceitual apresentado na Figura 3.

**Figura 3** - Modelo conceitual da entidade definidas no esquema do protocolo CTD



Fonte: Elaboração dos autores (2025).

A Figura 3 apresenta um diagrama de entidade-relacionamento (ER), que representa a modelagem conceitual dos relacionamentos entre entidades definidas a partir dos conjuntos de dados coletados na Espacenet. Assim, este modelo é um guia para a construção do esquema lógico e a subsequente criação do banco de dados relacional.

Mediante o exposto, foi utilizada a linguagem de programação Python para implementação de todas as estratégias propostas neste estudo. A escolha da linguagem de programação Python para o desenvolvimento da solução justifica-se por sua ampla aceitação e reconhecimento, especialmente pela simplicidade e eficiência que oferece no desenvolvimento de ferramentas de mineração e análise de dados. Além disso, Python é, atualmente (2025), uma das linguagens mais utilizadas no setor tecnológico, o que reforça sua adequação para este propósito.

## RESULTADOS

Os resultados apresentados neste estudo demonstram a eficácia da estratégia proposta em coletar e armazenar dados bibliográficos de patentes. Todas as estratégias foram implementadas por meio de uma combinação de técnicas de web scraping, organização e recuperação de dados e a linguagem de programação Python. Esse esforço resultou no desenvolvimento de um conjunto de scripts que caracterizam um framework de automação de coleta de dados bibliográficos de patentes em repositórios de acesso aberto.

Neste estudo, o *framework* estabelecido foi executado com o objetivo de coletar patentes depositadas no Brasil dentro do período de 1900 a 2024, visando recuperar um conjunto amplo e representativo de dados. Usando o repositório Espacenet como fonte de dados.

A partir dessa abordagem, foi possível recuperar um total de 946.367 documentos de patentes depositadas no Brasil, dos quais 831.706 correspondem a registros únicos, os demais registros são atualizações de patentes já depositadas.

Em relação ao tempo necessário para a coleta, o protocolo CTDP leva, em média, 2,5 segundos para pesquisar e recuperar os dados bibliográficos de uma patente no repositório Espacenet. Esse intervalo é necessário para atender às diretrizes de coleta e evitar bloqueios no acesso. Assim, para recuperar os 946.367 documentos de patentes, seriam necessários 2.365.918 segundos, o equivalente a aproximadamente 55 meses de execução contínua.

Com o objetivo de reduzir significativamente esse tempo, o protocolo foi executado em sete instâncias simultaneamente, cada uma responsável por um intervalo específico dentro do período de 1900 a 2024. Essa estratégia permitiu a coleta integral do conjunto de dados em apenas 7 meses, demonstrando a eficácia de adoção de uma abordagem paralela.

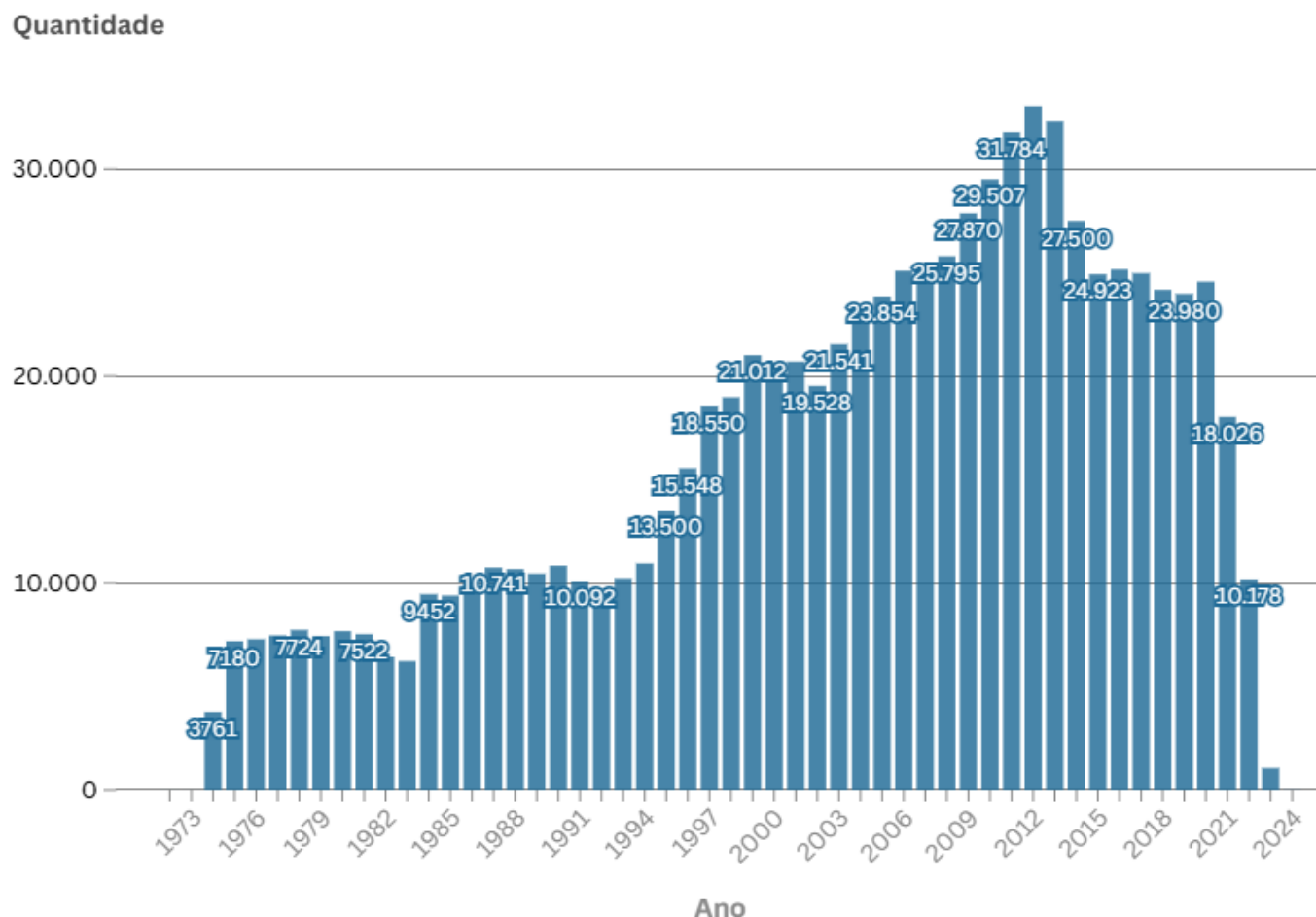
Considerando o volume de patentes a ser coletado em futuros estudos, recomenda-se a utilização de múltiplas instâncias do protocolo CTDP para otimizar o tempo de coleta. Além disso, é aconselhável priorizar a execução fora do horário comercial europeu, período em que, geralmente, os tempos de resposta são mais curtos, contribuindo para uma coleta mais eficiente.

O *framework* foi desenvolvido para ser compatível com os sistemas operacionais Windows, Linux e macOS, exigindo apenas que o usuário tenha previamente instalado o Python, na versão 3.10 ou superior. Em relação aos requisitos de hardware, recomenda-se um mínimo de 4 GB de memória RAM disponível, enquanto o espaço em disco necessário pode variar conforme o volume de dados a ser coletado.

Ademais, com objetivo de explorar o conjunto de dados coletados. Foi realizada uma análise dos depósitos anuais de patentes no período compreendido entre 19702 (a patente mais antiga coletada na Espacenet é de 1972) e 2024 (Ano limite de análise). Essa análise baseia-se no conjunto de patentes únicas. Os resultados dessa análise são apresentados na Figura 2, que apresenta a evolução temporal do total de depósitos de patentes. É importante

destacar que registros de patentes que não continham a informação do ano de depósito foram excluídos da análise, a fim de assegurar a consistência dos dados apresentados.

**Figura 4** - Evolução temporal do depósito de patentes por ano.



Fonte: Elaboração dos autores (2025).

O gráfico apresentado evidencia um crescimento expressivo no número de depósitos de patentes no Brasil ao longo dos anos, refletindo uma intensificação da atividade inovadora e do interesse na proteção de ideias e invenções no país. Durante o período analisado, a média anual de depósitos foi de aproximadamente 16.000 patentes. Ao restringir a análise aos últimos dez anos do período considerado, essa média aumenta para 17.403,60 depósitos anuais, destacando os anos de 2012, 2013 e 2011 como os mais produtivos, com 33.040, 32.345 e 31.784 depósitos, respectivamente.

A partir da década de 1990, observa-se um crescimento mais consistente e acelerado nos depósitos de patentes. Que sugere um período de fortalecimento do ambiente de negócios e da cultura de inovação no Brasil, sugerindo um período de avanço tecnológico. Esse cres-

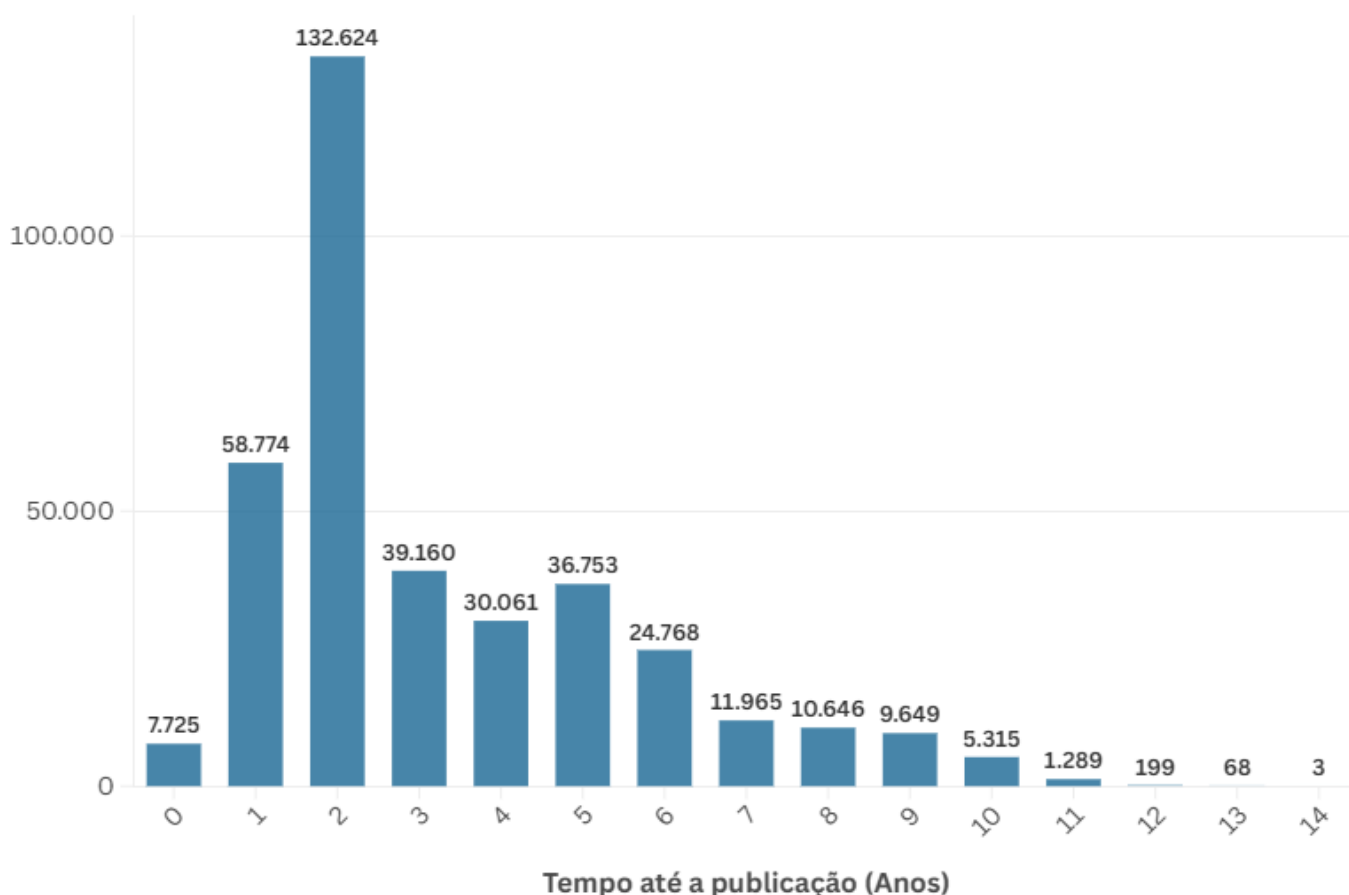
cimento pode estar associado a uma maior conscientização sobre a relevância da propriedade intelectual e interesse de empresas internacionais em protegerem seus inventos no Brasil.

No entanto, é perceptível uma redução acentuada no número de depósitos a partir de 2020. Esse declínio não condiz com os indicadores apresentados pelo INPI em maio de 2024, que apontam uma média de 27.267,50 depósitos anuais entre 2020 e 2024 (INPI, 2024). Essa discrepância sugere a possibilidade de que uma parcela dos depósitos realizados nesse período não estivesse disponível na base da Espacenet no momento da coleta de dados. Tal lacuna pode ser explicada por processos burocráticos que atrasam a disponibilização dos registros na Espacenet ou pelo fato de estarem em período de sigilo.

Com o objetivo de sustentar a hipótese anteriormente apresentada, a Figura 3 apresenta o intervalo de tempo transcorrido entre o depósito de uma patente e sua publicação. O cálculo baseia-se na diferença entre o ano de publicação e o ano de depósito.

**Figura 5 - Tempo de depósito até a publicação de patentes no período de 2010 a 2024.**

#### Quantidade de patentes



Fonte: Elaboração dos autores (2025).



A análise evidencia que a maioria das patentes depositadas no Brasil é publicada em um intervalo de dois anos após o depósito. Entretanto, a média geral do tempo de publicação é de sete anos, indicando variações consideráveis no processamento. Além disso, identificaram-se casos extremos em que a publicação de uma patente levou até 14 anos desde o depósito.

Esses resultados reforçam a hipótese de que muitas patentes depositadas durante o período de análise podem não estar disponíveis na base de dados utilizada, em função de atrasos na publicação ou restrições temporárias de acesso. Essa limitação impacta diretamente os resultados, subestimando o total de patentes depositadas nos últimos 5 anos.

## CONCLUSÃO

Os repositórios de dados de patentes constituem ferramentas fundamentais no processo de pesquisa e obtenção de documentos patentários, sendo amplamente empregados na prospecção tecnológica para identificar possibilidades de exploração ou aprimoramento de tecnologias e produtos, além de desempenharem um papel relevante no ambiente acadêmico. Entretanto, a coleta e organização desses dados não são tarefas triviais. Os repositórios de acesso aberto enfrentam limitações em relação ao volume de dados que podem manipular, o que, por conseguinte, dificulta a manipulação de grandes conjuntos de dados.

Por meio desta pesquisa, foram examinadas as dificuldades inerentes à manipulação de extensos conjuntos de dados provenientes de patentes utilizando bases de dados de acesso aberto, identificando-se as suas limitações e sugerindo abordagens eficazes para superar tais desafios. Nesse sentido, as estratégias propostas neste estudo se mostraram uma medida promissora, permitindo alcançar resultados satisfatórios, especialmente ao viabilizar análises que demandam grandes volumes de dados. Assim, essa abordagem emerge como uma estratégia viável para a recuperação de grandes conjuntos de documentos de patentes.

Ademais, ao empregar esta abordagem sistemática, espera-se que este estudo possa fornecer uma contribuição substancial para a ampliação do escopo dos estudos relacionados a documentos de patentes, proporcionando novas perspectivas e possibilidades para o progresso em diversas áreas do conhecimento.

## BIBLIOGRAFIA

Abrantes, P. C., Barros Junior, D., Silva, E. A. da, Islabão, G. I., & Santos, D. A. (2022). Estudo comparativo de bases gratuitas de patentes: Patentscope (WIPO), Espacenet (EPO), BuscaWeb (INPI/BR). *Revista Tecnologia e Sociedade*, 18(54), 125–142. <https://revistas.utfpr.edu.br/rtss/article/viewFile/14516/9108>





- Brandão, F. G. (2016). *Democratização da informação a partir do uso de repositórios digitais institucionais: da comunicação científica às informações tecnológicas de patentes* (Dissertação de mestrado, Universidade Regional Integrada do Alto Uruguai e das Missões) Repositório Lume. <https://lume.ufrgs.br/handle/10183/179853>
- Date, C. J. (2003). *Introdução a sistemas de banco de dados* (8a ed.). Elsevier.
- Espacenet. (2025). *Espacenet patent search*. <https://worldwide.espacenet.com/patent/>
- European Patent Office. (2025). *Open Patent Services RESTful Web Services* (Versão 1.3.20). <https://www.epo.org/searching-for-patents/data/web-services/ops.html>
- Google. (2025). *About Google Patents overview*. Google Patents. [https://support.google.com/faqs/answer/6390996?hl=pt-BR&ref\\_topic=6390989&sjid=4104479748124394271-SA](https://support.google.com/faqs/answer/6390996?hl=pt-BR&ref_topic=6390989&sjid=4104479748124394271-SA)
- Instituto Nacional da Propriedade Industrial. (2024, maio 17). *Pedidos de patentes de residentes crescem 10% em 2023*. Gov.br. <https://www.gov.br/inpi/pt-br/central-de-conteudo/noticias/pedidos-de-patentes-de-residentes-crescem-10-em-2023>
- Latipat. (2025). *Latipat-Espacenet*. Latipat-Espacenet. <https://lp.espacenet.com/>
- Lei nº 9.279, de 14 de maio de 1996* (1996, 14 de maio). Regula direitos e obrigações relativos à propriedade industrial. Presidência da República. [http://www.planalto.gov.br/ccivil\\_03/leis/19279.htm](http://www.planalto.gov.br/ccivil_03/leis/19279.htm)
- Nascimento, M. G. R. da S., & Speziali. (2020, novembro). Patentometria: A utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos NITs. In *Anais do IV Encontro Internacional de Gestão, Desenvolvimento e Inovação*. Edição online. <https://periodicos.ufms.br/index.php/EIGEDIN/article/view/11705/8252>
- Pires, E. A., Ribeiro, N. M., & Quintella, C. M. (2020, março). Sistemas de busca de patentes: Análise comparativa entre Espacenet, Patentscope, Google Patents, Lens, Derwent Innovation Index e Orbit Intelligence. *Cadernos de Prospecção*, 13(1), 13, 13-29. <https://periodicos.ufba.br/index.php/nit/article/view/35147>
- Questel. (2025). *Emerge smarter with Orbit Intelligence! Meet the future with agility & confidence*. Questel. <https://www.questel.com/patent/ip-intelligence-software/orbit-intelligence/>
- Reymond, D., & Quoniam, L. (2018). Patent documents in STEM and PhD education: Open-source tools and some examples to open discussion. In *IEEE Global Engineering Education Conference (EDUCON)* (pp. 4–9). IEEE. <https://doi.org/10.1109/EDUCON.2018.8363203>
- Rezende, N. G., Dalip, D. H., Brandão, M. A., & Vasconcelos, M. A. (2023). Elaboração de um conjunto de dados sobre o registro de patentes no Brasil. In *Anais do V Dataset Showcase Workshop (DSW)* (pp. 99–108). Sociedade Brasileira de Computação. <https://sol.sbc.org.br/index.php/dsw/article/view/25509/25329>



Sanz-Casado, E. (2006). *Los estudios métricos de la información y la evaluación de la actividad científica: conceptos básicos*. [Material didático] ECA/USP.

Singh, V., Chakraborty, K., & Lavina-Vincent, C. (2016, janeiro). Patent database: Their importance in prior art documentation and patent search. *Journal of Intellectual Property Rights*, 21(1), 42-56.

The SQLite Consortium. (2023). *The SQLite Consortium*. <https://sqlite.org/index.html>

The Lens. (2025). *About The Lens*. Cambia. <https://about.lens.org/#>

## ANEXO 1

### RESUMO BIOGRÁFICO DOS AUTORES

#### Raulivan Rodrigo da Silva

Doutorando em Modelagem Matemática e Computacional pelo CEFET-MG (2022). Mestre em Modelagem Matemática e Computacional pelo CEFET-MG (2022). Atualmente sou professor efetivo do CEFET MG no campus Divinópolis, lotado no Departamento de Informática, Gestão e Design (2019). ORCID: 0000-0002-2740-1045

#### Thiago Magela Rodrigues Dias

Doutor em Modelagem Matemática e Computacional pelo CEFET-MG (2016) tendo trabalhado com Bibliometria, Extração de Dados Científicos e Análise de Redes de Colaboração Científica. Mestre em Modelagem Matemática e Computacional pelo CEFET-MG (2008). Possui graduação em Ciência da Computação pelo Centro Universitário de Formiga - UNIFOR (2004), além de Especialização em Produção de Software - com Ênfase em Software Livre pela UFLA (2007) e Especialização em Melhoria do Processo de Software, UFLA (2007). Atua como Professor no Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG). ORCID: 0000-0002-3090-9413

#### Washington Luís Ribeiro de Carvalho Segundo

É Doutor e Mestre em Informática pela Universidade de Brasília, com Estágio de Doutorado Sanduíche no King's College London. Possui graduação em Matemática (Bacharelado e Licenciatura) também pela Universidade de Brasília. É Coordenador Técnico da Área de Tratamento, Análise e Disseminação da Informação Científica no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict / MCTI). ORCID: 0000-0003-3635-9384