



CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS  
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

# **UM PROTOCOLO SISTEMÁTICO PARA COLETA E TRATAMENTO DE DADOS PATENTÁRIOS EM REPOSITÓRIOS DE ACESSO ABERTO**

**RAULIVAN RODRIGO DA SILVA**

Orientador: Prof. Dr. Thiago Magela Rodrigues Dias  
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Washington Luís Ribeiro de Carvalho Segundo  
Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT, Brasil

BELO HORIZONTE  
DEZEMBRO DE 2025

**RAULIVAN RODRIGO DA SILVA**

# **UM PROTOCOLO SISTEMÁTICO PARA COLETA E TRATAMENTO DE DADOS PATENTÁRIOS EM REPOSITÓRIOS DE ACESSO ABERTO**

Tese apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Sistemas Inteligentes

Orientador: Prof. Dr. Thiago Magela Rodrigues Dias  
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Washington Luís Ribeiro de Carvalho Segundo  
Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT, Brasil

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS  
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL  
BELO HORIZONTE  
DEZEMBRO DE 2025

S586p Silva, Raulivan Rodrigo da  
Um protocolo sistemático para coleta e tratamento de dados patentários em repositórios de acesso aberto / Raulivan Rodrigo da Silva. – 2025.  
196 f.

Tese (doutorado) apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.  
Orientador: Thiago Magela Rodrigues Dias.  
Coorientador: Washington Luís Ribeiro de Carvalho Segundo.  
Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Patentes – Teses. 2. Pesquisa industrial – Legislação – Teses. 3. Inovações tecnológicas – Teses. 4. Patentes – Idioma – Teses. 5. Propriedade intelectual – Estatísticas – Teses. I. Dias, Thiago Magela Rodrigues. II. Carvalho Segundo, Washington Luís Ribeiro de. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 001.422



SERVIÇO PÚBLICO FEDERAL  
MINISTÉRIO DA EDUCAÇÃO  
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS  
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

## **“UM PROTOCOLO SISTEMÁTICO PARA COLETA E TRATAMENTO DE DADOS PATENTÁRIOS EM REPOSITÓRIOS DE ACESSO ABERTO”.**

Tese de Doutorado apresentada por **Raulivan Rodrigo da Silva**, em 17 de dezembro de 2025, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:



Documento assinado digitalmente  
**THIAGO MAGELA RODRIGUES DIAS**  
Data: 03/02/2026 17:08:13-0300  
Verifique em <https://validar.iti.gov.br>

**Prof. Dr. Thiago Magela Rodrigues Dias (Orientador)**  
Centro Federal de Educação Tecnológica de Minas Gerais



Documento assinado digitalmente  
**WASHINGTON LUIS RIBEIRO DE CARVALHO SEGUNDO**  
Data: 16/02/2026 10:07:47-0300  
Verifique em <https://validar.iti.gov.br>

**Prof. Dr. Washington Luís Ribeiro de Carvalho Segundo (Coorientador)**  
Instituto Brasileiro de Informação em Ciência e Tecnologia

**Prof. Dr. Fabio Lorensi do Canto**  
Universidade Federal de Santa Catarina



Documento assinado digitalmente  
**FABIO LORENSI DO CANTO**  
Data: 05/02/2026 19:09:21-0300  
Verifique em <https://validar.iti.gov.br>

**Profª. Drª. Thabatta Moreira Alves de Araújo**  
Centro Federal de Educação Tecnológica de Minas Gerais



Documento assinado digitalmente  
**THABATTA MOREIRA ALVES DE ARAUJO**  
Data: 06/02/2026 13:02:22-0300  
Verifique em <https://validar.iti.gov.br>

**Prof. Dr. Gray Farias Moita**  
Centro Federal de Educação Tecnológica de Minas Gerais



Documento assinado digitalmente  
**GRAY FARIAS MOITA**  
Data: 05/02/2026 16:04:31-0300  
Verifique em <https://validar.iti.gov.br>

**Prof. Dr. Gustavo Campos Menezes**  
Centro Federal de Educação Tecnológica de Minas Gerais



Documento assinado digitalmente  
**GUSTAVO CAMPOS MENEZES**  
Data: 06/02/2026 11:29:14-0300  
Verifique em <https://validar.iti.gov.br>

Visto e permitida à impressão,

**Prof. Dr. Alisson Marques da Silva**  
Coordenador do Programa de Pós-Graduação Stricto Sensu em  
Modelagem Matemática e Computacional

Dedico este trabalho aos meus pais, minha esposa, minha irmã, meus filhos e aqueles(as) que estiveram ao meu lado em toda esta trajetória.

# Agradecimentos

Agradeço primeiramente a Deus por ter me abençoado durante este estudo com saúde, força e sabedoria para conseguir chegar aqui.

Aos meus pais, minha irmã, minha esposa e meus filhos, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste estudo.

A todos meus professores, que compartilharam seus conhecimentos, em especial ao meu orientador, Dr. Thiago Magela Rodrigues Dias e meu coorientador Dr. Washington Luís Ribeiro de Carvalho Segundo, por acreditar e incentivar o desenvolvimento deste estudo, apoiando e orientando com dedicação e amizade.

Ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET/MG), pelo suporte e incentivo em minha formação acadêmica, por proporcionar toda estrutura adequada para que eu pudesse executar as minhas atividades.

A todos que participaram, direta ou indiretamente do desenvolvimento deste estudo, contribuindo e enriquecendo o meu processo de aprendizado.

*“O princípio fundamental do saber é: procura  
obter sabedoria; investe todo o teu ser e o  
que possuis para adquirir entendimento.”  
(Provérbios 4: 7)*

# Resumo

Este estudo busca contribuir para o avanço e a disseminação de pesquisas na área de Propriedade Intelectual, com ênfase em dados bibliográficos de patentes. Através de uma abordagem multidisciplinar, que relaciona disciplinas da Ciência da Computação e da Ciência da Informação, este estudo tem como objetivo principal propor um protocolo sistemático de coleta e tratamento de grandes conjuntos de dados bibliográficos de patentes, visando superar as limitações inerentes aos repositórios de acesso aberto a tais documentos. A fundamentação deste estudo é sustentada pela elaboração de um conjunto de estratégias voltadas à construção de um protocolo, compreendido como um conjunto de procedimentos adotados para a coleta e o processamento de dados bibliográficos de patentes. Esse protocolo combina técnicas de organização e recuperação da informação, inteligência artificial e patentometria. Para compreender os desafios envolvidos na recuperação de dados bibliográficos de patentes em repositórios de acesso aberto, foi realizada uma pesquisa aplicada, exploratória e dedutiva que resultou no estabelecimento de onze critérios que orientaram o pesquisador no processo de seleção dos repositórios utilizados como fontes de dados. A partir dessa seleção, foram delineadas estratégias para viabilizar a coleta, o tratamento e a agregação dos conjuntos de dados, resultando na construção de uma base de dados local capaz de subsidiar análises sob diferentes perspectivas. Além disso, com o intuito de enriquecer essa base de dados, foi elaborado um catálogo de fontes complementares, composto por relatórios disponibilizados pelo Instituto Nacional da Propriedade Industrial (INPI) e por registros da base curricular da Plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que fornecem informações sobre os pesquisadores que informaram patentes em seus currículos. Para validar a eficácia do protocolo proposto, ele foi aplicado na recuperação de patentes depositadas no Brasil entre 1900 e 2024, usando como fonte de dados o repositório Espacenet. Os dados coletados foram submetidos a uma análise detalhada sob diferentes perspectivas, utilizando técnicas de bibliometria, patentometria, mineração de texto e processamento de linguagem natural. Os resultados obtidos demonstram a relevância do protocolo proposto neste estudo, evidenciando sua aplicabilidade em pesquisas que envolvem a análise de grandes conjuntos de documentos de patentes e oferecendo uma alternativa relevante para superar os desafios inerentes à coleta de dados em repositórios de acesso aberto.

**Palavras-chave:** Patentes, Protocolo, Espacenet, INPI, Patentometria, Propriedade Intelectual.

# Abstract

This study aims to contribute to the advancement and dissemination of research in the field of Intellectual Property, with an emphasis on bibliographic patent data. Through a multidisciplinary approach, relating disciplines of Computer Science and Information Science, this study aims to propose a systematic protocol for collecting and processing large sets of bibliographic patent data, aiming to overcome the limitations inherent in open access repositories for such documents. The foundation of this study is supported by the development of a set of strategies aimed at constructing a protocol, understood as a set of procedures adopted for the collection and processing of bibliographic patent data. This protocol combines information organization and retrieval techniques, artificial intelligence, and patentometrics. To understand the challenges involved in retrieving bibliographic patent data from open access repositories, an applied, exploratory, and deductive research was conducted, resulting in the establishment of eleven criteria that guided the researcher in the process of selecting the repositories used as data sources. Based on this selection, strategies were outlined to enable the collection, processing, and aggregation of the datasets, resulting in the construction of a local database capable of supporting analyses from different perspectives. Furthermore, in order to enrich this database, a catalog of complementary sources was created, composed of reports made available by the National Institute of Industrial Property (INPI) and records from the Lattes Platform curriculum database, maintained by the National Council for Scientific and Technological Development (CNPq), which provide information on researchers who reported patents in their curricula. To validate the effectiveness of the proposed protocol, it was applied to the retrieval of patents filed in Brazil between 1900 and 2024, using the Espacenet repository as a data source. The collected data were subjected to a detailed analysis from different perspectives, using techniques of bibliometrics, patentometrics, text mining, and natural language processing. The results obtained demonstrate the relevance of the protocol proposed in this study, highlighting its applicability in research involving the analysis of large sets of patent documents and offering a relevant alternative to overcome the challenges inherent in data collection from open access repositories.

**Keywords:** Patents, Protocol, Espacenet, INPI, Patentometrics, Intellectual Property.

# Lista de Figuras

|                                                                                                                          |     |
|--------------------------------------------------------------------------------------------------------------------------|-----|
| Figura 1 – Ciclo de relacionamento entre criador e sociedade . . . . .                                                   | 9   |
| Figura 2 – Grupos da Propriedade Intelectual . . . . .                                                                   | 10  |
| Figura 3 – Detalhamento de uma classificação de patente . . . . .                                                        | 18  |
| Figura 4 – Resumo do sistema do PCT . . . . .                                                                            | 20  |
| Figura 5 – Pedido de patente US2008/0200653 A1 . . . . .                                                                 | 21  |
| Figura 6 – busca básica de patentes no INPI . . . . .                                                                    | 24  |
| Figura 7 – busca avançada de patentes no INPI . . . . .                                                                  | 25  |
| Figura 8 – Processo genérico para coleta e análise patentes . . . . .                                                    | 32  |
| Figura 9 – Linha do tempo da evolução dos estudos métricos. . . . .                                                      | 35  |
| Figura 10 – Diagrama de inter-relação entre os subcampos . . . . .                                                       | 36  |
| Figura 11 – Esquema do fluxo dos processos operacionais de pesquisa. . . . .                                             | 49  |
| Figura 12 – Fluxo operacional para a construção da metodologia de coleta e trata-<br>mento de dados de patentes. . . . . | 53  |
| Figura 13 – Visão geral do capítulo de desenvolvimento . . . . .                                                         | 57  |
| Figura 14 – Esquema de relacionamento entre os módulos . . . . .                                                         | 63  |
| Figura 15 – Protocolo CTDP no processo genérico . . . . .                                                                | 65  |
| Figura 16 – Representação do arquivo de patentes no formato Espacenet . . . . .                                          | 78  |
| Figura 17 – Modelo conceitual da entidade definidas no esquema do protocolo CTDP                                         | 80  |
| Figura 18 – Exemplo de construção de um <i>hash</i> de uma patente . . . . .                                             | 86  |
| Figura 19 – Repositório de dados selecionados para a construção da base de dados<br>local . . . . .                      | 89  |
| Figura 20 – Visão geral do processo de coleta de dados . . . . .                                                         | 90  |
| Figura 21 – Coleta de patentes na Espacenet . . . . .                                                                    | 91  |
| Figura 22 – Organização de arquivos de patente . . . . .                                                                 | 92  |
| Figura 23 – Visão geral do LattesDataXplorer . . . . .                                                                   | 94  |
| Figura 24 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo A                                        | 96  |
| Figura 25 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo B                                        | 96  |
| Figura 26 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo C                                        | 97  |
| Figura 27 – Formato número de depósito de patentes adotado pelo INPI até 12/2011                                         | 99  |
| Figura 28 – Formato atual de número de depósito de patentes adotado pelo INPI . .                                        | 100 |
| Figura 29 – Comparativo entre o número de depósito informado no INPI e na Espacenet                                      | 100 |
| Figura 30 – Fluxo de aplicação das regras . . . . .                                                                      | 103 |
| Figura 31 – Pipeline de <i>Machine Learning</i> e Processamento de Linguagem Natural .                                   | 105 |
| Figura 32 – Página inicial do Pytente . . . . .                                                                          | 113 |
| Figura 33 – Evolução temporal do depósito de patentes por ano . . . . .                                                  | 117 |
| Figura 34 – Tempo de depósito até a publicação de patentes no período de 2010 a 2024                                     | 119 |

|                                                                                                                                                                                       |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Figura 35 – Total de patentes pela a primeira classificação atribuída no processo de depósito . . . . .                                                                               | 120 |
| Figura 36 – Total de depósito de patentes por ano e por classificação . . . . .                                                                                                       | 122 |
| Figura 37 – Distribuição das classificações agrupadas pela seção e a classe que compõem a primeira classificação atribuída à patente no processo de depósito. . . . .                 | 124 |
| Figura 38 – Representação do quantitativo de classificações agrupadas pela seção e a classe que compõem a primeira classificação atribuída à patente no processo de depósito. . . . . | 126 |
| Figura 39 – Termos mais frequentes nos títulos das patentes . . . . .                                                                                                                 | 128 |
| Figura 40 – Agrupamento de patentes por termos semelhantes . . . . .                                                                                                                  | 129 |
| Figura 41 – Matriz de similaridade entre os tópicos . . . . .                                                                                                                         | 132 |
| Figura 42 – Rede de relacionamento estabelecida entre os termos que caracterizam os principais tópicos identificando pela análise . . . . .                                           | 133 |
| Figura 43 – Indicador de modularidade de classe da rede de relacionamento de termos                                                                                                   | 134 |
| Figura 44 – Árvore de hierarquia de tópicos . . . . .                                                                                                                                 | 135 |
| Figura 45 – Total de depositantes por nacionalidade informada no processo de depósito de patente . . . . .                                                                            | 139 |
| Figura 46 – Patentes certificadas e patentes não certificadas informadas no currículos da Plataforma Lattes . . . . .                                                                 | 141 |
| Figura 47 – Patentes certificadas e patentes não certificadas por ano de depósito informado nos currículos da Plataforma Lattes . . . . .                                             | 142 |
| Figura 48 – Formação dos proponentes de patentes . . . . .                                                                                                                            | 145 |
| Figura 49 – Áreas de atuação dos proponentes de patentes . . . . .                                                                                                                    | 146 |
| Figura 50 – Distribuição de proponentes de patentes por estado . . . . .                                                                                                              | 147 |
| Figura 51 – Idade científica . . . . .                                                                                                                                                | 149 |
| Figura 52 – Total de depósito por idade científica . . . . .                                                                                                                          | 150 |
| Figura 53 – 30 IES depositantes de patentes . . . . .                                                                                                                                 | 152 |
| Figura 54 – Atualização dos currículos da Plataforma Lattes . . . . .                                                                                                                 | 153 |
| Figura 55 – Exemplo de rede de colaboração estabelecida entre três inventores . . .                                                                                                   | 155 |
| Figura 56 – Visão geral da rede de colaboração . . . . .                                                                                                                              | 156 |
| Figura 57 – Sub-rede de entre os pesquisadores mais colaborativos . . . . .                                                                                                           | 159 |
| Figura 58 – Sub-rede com as maiores comunidades identificadas . . . . .                                                                                                               | 161 |
| Figura 59 – Página do guia passo a passo de instalação do Pytente . . . . .                                                                                                           | 189 |
| Figura 60 – Página do guia passo a passo de instalação do Pytente . . . . .                                                                                                           | 191 |
| Figura 61 – Página de coleta de dados no repositório Espacenet . . . . .                                                                                                              | 192 |
| Figura 62 – Tela de sucesso de execução do processo de coleta de dados bibliográficos de patentes no repositório Espacenet . . . . .                                                  | 193 |
| Figura 63 – Tela de configuração de carga de dados bibliográficos de patentes . . .                                                                                                   | 194 |

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| Figura 64 – Tela de conclusão do processo de carga dos dados . . . . . | 195 |
|------------------------------------------------------------------------|-----|

# Lista de Tabelas

|                                                                                                                       |     |
|-----------------------------------------------------------------------------------------------------------------------|-----|
| Tabela 1 – Resumo das principais características dos repositórios de patentes internacionais . . . . .                | 30  |
| Tabela 2 – Restrições inerentes à recuperação de dados de patentes nos repositórios analisados neste estudo . . . . . | 60  |
| Tabela 3 – Matriz de Triagem de Repositórios de Patentes . . . . .                                                    | 70  |
| Tabela 4 – Tabelas BADEPI . . . . .                                                                                   | 76  |
| Tabela 5 – Listagem de pastas . . . . .                                                                               | 93  |
| Tabela 7 – Similaridade de títulos usando Levenshtein . . . . .                                                       | 104 |
| Tabela 8 – Quantitativos de patentes coletadas no repositório Espacenet . . . . .                                     | 114 |
| Tabela 10 – Quantitativos de registros da Plataforma Lattes . . . . .                                                 | 115 |
| Tabela 12 – Palavras-chave que compõem os dez tópicos mais relevantes . . . . .                                       | 131 |
| Tabela 13 – Amostragem de resultados da busca semântica . . . . .                                                     | 136 |
| Tabela 14 – Amostragem de resultados menos relevantes da busca semântica . . . . .                                    | 137 |
| Tabela 15 – Detalhamento da certificação . . . . .                                                                    | 144 |
| Tabela 16 – Estatísticas da rede de colaboração entre os pesquisadores . . . . .                                      | 157 |
| Tabela 18 – Resultados considerados na análise . . . . .                                                              | 185 |

# Lista de Quadros

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Quadro 1 – Códigos INID . . . . .                                                                    | 22  |
| Quadro 2 – Amostragem de indicadores derivados dos código INDI . . . . .                             | 38  |
| Quadro 3 – Indicadores patentométricos INPI . . . . .                                                | 39  |
| Quadro 4 – Referencial metodológico adotado na tese . . . . .                                        | 48  |
| Quadro 5 – Compilado das fontes de pesquisas adotadas na tese. . . . .                               | 51  |
| Quadro 6 – Grupo (02) ferramentas de análise e/ou de busca de patentes . . . . .                     | 59  |
| Quadro 7 – Estratégias que objetivam a solução dos desafios encontrados. . . . .                     | 61  |
| Quadro 8 – Critérios de seleção . . . . .                                                            | 67  |
| Quadro 9 – Entidade Patente . . . . .                                                                | 80  |
| Quadro 10 – Entidade Classificação . . . . .                                                         | 81  |
| Quadro 11 – Entidade Pessoa . . . . .                                                                | 81  |
| Quadro 12 – Entidades Depositante, Inventor e Procurador . . . . .                                   | 82  |
| Quadro 13 – Entidades Identificador . . . . .                                                        | 82  |
| Quadro 14 – Entidade Indivíduo Lattes . . . . .                                                      | 83  |
| Quadro 15 – Entidade Patente Lattes . . . . .                                                        | 83  |
| Quadro 16 – Critérios de validação . . . . .                                                         | 84  |
| Quadro 17 – Critérios de deduplicação . . . . .                                                      | 85  |
| Quadro 18 – Descrição das tabelas que compõem banco de dados local . . . . .                         | 95  |
| Quadro 19 – Códigos de natureza de proteção que compõem o número de depósito<br>da patente . . . . . | 188 |
| Quadro 20 – Códigos de publicação que compõem o número de depósito da patente                        | 188 |

# Lista de Algoritmos

|                                                                     |    |
|---------------------------------------------------------------------|----|
| Algoritmo 1 – PESQUISA DE PATENTES NA ESPACENET . . . . .           | 74 |
| Algoritmo 2 – COLETA DE DADOS BIBLIOGRÁFICOS NA ESPACENET . . . . . | 75 |

# Lista de Abreviaturas e Siglas

|          |                                                                    |
|----------|--------------------------------------------------------------------|
| API      | <i>Application Programming Interface</i>                           |
| BADEPI   | Base de Dados sobre Propriedade Intelectual para Fins Estatísticos |
| BERT     | <i>Bidirectional Encoder Representations from Transformers</i>     |
| CAPES    | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior        |
| CEFET-MG | Centro Federal de Educação Tecnológica de Minas Gerais             |
| CIP      | Classificação Internacional de Patente                             |
| CI       | Ciência da Informação                                              |
| CNPq     | Conselho Nacional de Desenvolvimento Científico e Tecnológico      |
| CNPJ     | Cadastro Nacional da Pessoa Jurídica                               |
| CPC      | <i>Cooperative Patent Classification</i>                           |
| CPF      | Cadastro de Pessoa Física                                          |
| CQL      | <i>Contextual Query Language</i>                                   |
| CSV      | <i>comma-separated-values</i>                                      |
| CTDP     | Coleta e Tratamento de Dados de Patentes                           |
| C&T      | Ciência e Tecnologia                                               |
| CUP      | Convenção da União de Paris                                        |
| DBLP     | <i>Digital Bibliography &amp; Library Project</i>                  |
| DII      | <i>Derwent Innovation Index</i>                                    |
| DOI      | <i>Digital Object Identifier</i>                                   |
| DOU      | Diário Oficial da União                                            |
| DSR      | <i>Design Science Research</i>                                     |
| DWPI     | <i>Derwent World Patents Index</i>                                 |
| ECLA     | <i>European Classification of Languages</i>                        |
| EPO      | <i>European Patent Office</i>                                      |

|         |                                                                      |
|---------|----------------------------------------------------------------------|
| ER      | Entidade Relacionamento                                              |
| ETL     | Extração, Transformação e Carga                                      |
| EUA     | Estados Unidos da América                                            |
| FIOCRUZ | Fundação Oswaldo Cruz                                                |
| FORTEC  | Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia |
| Gb      | <i>Gigabit</i>                                                       |
| GRU     | Guia de Recolhimento da União                                        |
| HTML    | <i>HyperText Markup Language</i>                                     |
| IBICT   | Instituto Brasileiro de Informação em Ciência e Tecnologia           |
| ICT     | Instituições de Ciência e Tecnologia                                 |
| IES     | Instituições de Ensino Superior                                      |
| INID    | <i>International Agreed Numbers for the Identification of Data</i>   |
| INPI    | Instituto Nacional de Propriedade Industrial                         |
| IPC     | <i>International Patent Classification</i>                           |
| IPI     | <i>Intellectual Property India</i>                                   |
| JPO     | <i>Japan Patent Office</i>                                           |
| JSON    | <i>JavaScript Object Notation</i>                                    |
| KIPO    | <i>Korean Intellectual Property Office</i>                           |
| LLM     | <i>Large Language Models</i>                                         |
| MEC     | Ministério da Educação                                               |
| MDIC    | Ministério do Desenvolvimento, Indústria e Comércio Exterior         |
| MD5     | <i>Message-Digest algorithm 5</i>                                    |
| MU      | Patente de Modelo de Utilidade                                       |
| NIT     | Núcleo de Inovação Tecnológica                                       |
| OEPM    | <i>Oficina Española de Patentes y Marcas</i>                         |
| OCED    | Organização para a Cooperação e Desenvolvimento Econômico            |

|         |                                                                   |
|---------|-------------------------------------------------------------------|
| OMPI    | Organização Mundial de Propriedade Intelectual                    |
| OPS     | <i>Open Patent Services</i>                                       |
| PCT     | Tratado de Cooperação de Patentes                                 |
| PDF     | <i>Portable Document Format</i>                                   |
| PI      | Patente de Invenção                                               |
| pePI    | Pesquisa em Propriedade Industrial                                |
| PNL     | Processamento de Linguagem Natural                                |
| PPGMMC  | Programa de Pós-Graduação em Modelagem Matemática e Computacional |
| P&D     | Pesquisa e Desenvolvimento                                        |
| P2N     | Patent2Net                                                        |
| RSLP    | Protocolo de Revisão Sistemática de Literatura Patentária         |
| SHA     | <i>Secure Hash Algorithm</i>                                      |
| SGBD    | Sistema Gerenciador de Banco de Dados                             |
| SGBDOR  | Sistema Gerenciador de Banco de Dados Objeto-Relacional           |
| SIGPPI  | Sistema Integrado de Gestão de Propriedade Intelectual            |
| SIPO    | <i>State Intellectual Property Office of China</i>                |
| SQL     | <i>Structured Query Language</i>                                  |
| TRCAPP  | Treinamento Rítmico Corporal Aplicativo                           |
| UFMG    | Universidade Federal de Minas Gerais                              |
| UFRGS   | Universidade Federal do Rio Grande do Sul                         |
| URL     | <i>Uniform Resource Locator</i>                                   |
| UNICAMP | Universidade Estadual de Campinas                                 |
| USP     | Universidade de São Paulo                                         |
| USPC    | <i>United States Patent Classification</i>                        |
| USPTO   | <i>United States Patent and Trademark Office</i>                  |

|      |                                                 |
|------|-------------------------------------------------|
| XML  | <i>eXtensible Markup Language</i>               |
| WEB  | <i>World Wide Web</i>                           |
| WIPO | <i>World Intellectual Property Organization</i> |
| WWW  | <i>World Wide Web</i>                           |

# Sumário

|                                                                                    |           |
|------------------------------------------------------------------------------------|-----------|
| <b>1 – Introdução</b>                                                              | <b>1</b>  |
| 1.1 Descrição do problema                                                          | 2         |
| 1.2 Justificativa                                                                  | 4         |
| 1.3 Objetivos                                                                      | 5         |
| 1.4 Público alvo e contribuições esperadas                                         | 5         |
| 1.5 Organização do trabalho                                                        | 6         |
| <b>2 – Fundamentação Teórica</b>                                                   | <b>8</b>  |
| 2.1 Propriedade Intelectual                                                        | 8         |
| 2.2 Patentes                                                                       | 11        |
| 2.2.1 Sistemas de classificação de patentes                                        | 14        |
| 2.2.2 Famílias de patentes                                                         | 18        |
| 2.2.3 Dados bibliográficos da patente                                              | 20        |
| 2.3 Repositórios de documentos de patentes                                         | 23        |
| 2.4 Coleta de dados de patentes                                                    | 31        |
| 2.5 Análise de patentes com ferramentas da Ciência da Informação                   | 34        |
| <b>3 – Trabalhos Relacionados</b>                                                  | <b>40</b> |
| 3.1 Patentes como objeto de pesquisa                                               | 40        |
| 3.2 Considerações sobre os trabalhos relacionados e diferenciais propostos         | 45        |
| <b>4 – Metodologia</b>                                                             | <b>47</b> |
| 4.1 Questão de pesquisa                                                            | 47        |
| 4.2 Hipótese                                                                       | 47        |
| 4.3 Estrutura referencial metodológica                                             | 47        |
| 4.4 Processos operacionais de pesquisa                                             | 49        |
| 4.4.1 Levantamento bibliográfico e documental                                      | 50        |
| 4.4.2 Construção da metodologia de coleta e tratamento dos dados de patentes       | 52        |
| <b>5 – Desenvolvimento</b>                                                         | <b>56</b> |
| 5.1 Avaliação das restrições existentes no processo de coleta de patentes          | 57        |
| 5.2 Protocolo de coleta e tratamento de dados bibliográficos de patentes           | 62        |
| 5.2.1 Triagem de repositório de dados de patentes e fontes de dados complementares | 66        |
| 5.2.2 Catálogo de fonte de dados complementares                                    | 71        |

|          |                                                                                                        |                |
|----------|--------------------------------------------------------------------------------------------------------|----------------|
| 5.2.3    | Coleta de dados                                                                                        | 72             |
| 5.2.4    | Limpeza e transformação dos dados                                                                      | 77             |
| 5.2.5    | Consulta e seleção de dados                                                                            | 86             |
| 5.3      | Estudo de praticabilidade e eficácia                                                                   | 87             |
| 5.3.1    | Pipeline para coleta e tratamento de dados de patentes                                                 | 88             |
| 5.4      | Certificação do conjunto de dados coletados                                                            | 98             |
| 5.5      | Pipeline de aprendizado de máquina e processamento de linguagem natural                                | 105            |
| <b>6</b> | <b>– Pytente</b>                                                                                       | <b>109</b>     |
| 6.1      | Estudo de praticabilidade e eficácia do protocolo CTD                                                  | 109            |
| 6.2      | Detalhamento técnico e procedimentos de uso do Pytente                                                 | 112            |
| <b>7</b> | <b>– Análise e Discussão dos Resultados</b>                                                            | <b>114</b>     |
| 7.1      | Apresentação e descrição dos dados                                                                     | 114            |
| 7.2      | Análise exploratória dos dados                                                                         | 116            |
| 7.3      | Certificação de patentes na Plataforma Lattes                                                          | 140            |
| 7.4      | Rede de colaboração estabelecida entre os pesquisadores com currículo registrados na Plataforma Lattes | 154            |
| <b>8</b> | <b>– Considerações Finais</b>                                                                          | <b>163</b>     |
| 8.1      | Conclusão e principais contribuições                                                                   | 163            |
| 8.2      | Trabalhos futuros                                                                                      | 167            |
| 8.3      | Publicações                                                                                            | 168            |
|          | <b>Referências</b>                                                                                     | <b>174</b>     |
|          | <br><b>Apêndices</b>                                                                                   | <br><b>184</b> |
|          | <b>APÊNDICE A – Resultados da busca no Google</b>                                                      | <b>185</b>     |
|          | <b>APÊNDICE B – Elementos da estrutura do número de depósito de patentes</b>                           | <b>188</b>     |
|          | <b>APÊNDICE C – Instalação e configuração do Pytente</b>                                               | <b>189</b>     |
| C.1      | Coleta de dados bibliográficos de patentes                                                             | 191            |
| C.2      | Tratamento e integração de dados                                                                       | 194            |
| C.3      | Extração de subconjuntos                                                                               | 196            |

# 1 Introdução

O século XXI tem sido um terreno fértil para o desenvolvimento de estruturas tecnológicas, com uma evolução cada vez mais acelerada. Novos dispositivos, aplicações e meios digitais surgem diariamente no mercado tecnológico, trazendo versões aprimoradas de funcionalidades já existentes ou apresentando soluções inovadoras. Diante desse cenário, as organizações têm se dedicado ao monitoramento contínuo de suas atividades, bem como à avaliação da viabilidade de seus produtos e serviços, a fim de implementar inovações que assegurem a fidelização e ampliação de sua base de clientes (AMADEI; TORKOMIAN, 2009).

Desde a formalização da propriedade intelectual, por meio de documentos de patentes, parte das inovações geradas têm origem em pesquisas realizadas em instituições de ensino públicas e privadas, cuja produção vem crescendo ao longo dos anos, consolidando-as como importantes pólos de inovação no cenário nacional. Esse desenvolvimento impulsionou a criação de leis e normas voltadas à proteção da propriedade intelectual gerada nas universidades, além de fomentar a criação de núcleos de apoio ao registro de patentes, promovendo a integração entre os setores acadêmico e empresarial. Isso tem incentivado instituições de ensino e pesquisadores a se envolverem cada vez mais no depósito de patentes (MUELLER; PERUCCHI, 2014).

Assim, 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, por sua vez, 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.

Diante do exposto, é coerente afirmar que os documentos de patentes constituem uma fonte rica e consistente de informações tecnológicas, uma vez que exigem descrições

minuciosas e detalhadas das invenções, incluindo desenhos técnicos, esquemas ilustrativos e reivindicações que delimitam o escopo de proteção da tecnologia apresentada. Esses documentos se destacam como recursos fundamentais para pesquisadores e estudiosos no campo da Propriedade Intelectual, pois possibilitam a compreensão aprofundada dos avanços tecnológicos registrados. Conforme [Reymond e Quoniam \(2018\)](#), os documentos de patentes apresentam uma quantidade de informações substancialmente superior àquela disponível em artigos científicos, ampliando sua relevância como instrumento de pesquisa, inovação e análise tecnológica.

De acordo com a Organização Mundial da Propriedade Intelectual (OMPI), cerca de 70% das informações divulgadas em documentos de patentes não estão disponíveis em outros tipos de documentos ([INPI, 2020](#)). Publicações da Organização para a Cooperação e Desenvolvimento Econômico (OCED) e da revista *Cadernos de Prospecção Tecnológica* e do Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia (FORTEC) evidenciam a importância de se utilizar documentos de patentes como fonte de informação tecnológica ([SILVA; NASCIMENTO, 2020](#)). Esses dados ressaltam a singularidade e a importância estratégica desses documentos no contexto da inovação e do desenvolvimento tecnológico, evidenciando o papel fundamental que as patentes desempenham na disseminação de conhecimento. É possível encontrar na literatura trabalhos, incluindo patentes, que apresentam técnicas para o gerenciamento dos dados que são publicados neste tipo de documento ([IBM \(2022\)](#); [NCET \(2019\)](#); [HQIT \(2017\)](#)).

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 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.

Na próxima seção estão delineados os problemas e desafios abordados neste estudo.

## 1.1 Descrição do problema

De acordo com dados da OMPI, foram realizados mundialmente 3,4 milhões de pedidos de patente em 2021, representando um crescimento de 3,6% em comparação aos

anos anteriores (WIPO, 2022). No Brasil, conforme informações do INPI, foram depositadas 20.343 patentes entre janeiro e setembro de 2024, o que representa um aumento de 1,5% em relação ao mesmo período do ano anterior (INPI, 2024a). Esses dados evidenciam uma tendência de crescimento no número de depósitos de patentes tanto em âmbito global quanto nacional.

Os documentos de patentes, tanto nacionais quanto estrangeiros, são disponibilizados para consulta pública em repositórios de patentes acessíveis pela internet, em conformidade com um dos princípios fundamentais do processo de depósito de patente: a divulgação pública do invento protegido. Praticamente todos os países que adotam a proteção por propriedade industrial possuem repositórios de patentes. No Brasil, o repositório nacional é mantido pelo INPI. Além disso, há repositórios internacionais, como a Espacenet, que permitem o acesso a documentos de patentes, assim como plataformas independentes de escritórios de patentes, como o Google Patents.

De fato, há diversos repositórios disponíveis na internet para consulta de dados de patentes, porém eles diferem em aspectos como os mecanismos de busca e as opções de exportação de dados. O acesso a documentos de patentes não é uma tarefa trivial, estando sujeito a diversas limitações, como o volume de dados que pode ser manipulado em um único processo, restrições de horário para consulta, além de limitações nos filtros e critérios de seleção. Essas barreiras dificultam a recuperação de conjuntos de dados mais amplos, como, por exemplo, a recuperação de todas as patentes depositadas em determinado país.

O repositório nacional mantido pelo INPI apresenta restrições significativas no que tange à consulta e à recuperação de grandes volumes de patentes. Conhecido como “pePI - Pesquisa em Propriedade Industrial” ou simplesmente “Busca Web”<sup>1</sup>, a interface deste repositório não é tão intuitiva quanto outras bases de dados internacionais, como a Espacenet. Embora ofereça algumas opções de filtros, como número do pedido, nome do depositante ou inventor e data de depósito, a flexibilidade na combinação desses critérios para refinar as buscas é limitada. Além disso, o sistema não permite a exportação em massa dos resultados de pesquisa. Não há recursos disponíveis para exportar os dados em formatos acessíveis à análise, como CSV, XML, JSON ou planilhas.

Diante do exposto, embora o conteúdo dos documentos de patentes seja de acesso público e gratuito, observa-se uma lacuna significativa entre a disponibilidade legal do dado e a sua acessibilidade analítica. Muitos pesquisadores frequentemente recorrem a repositórios privados, como o Derwent Innovation Index ou o Orbit Intelligence, não pelo conteúdo em si, mas pela infraestrutura de tratamento e exportação de dados que tais plataformas oferecem. Isso gera uma dualidade em relação ao direito da sociedade ao acesso público e gratuito aos documentos de patentes. O documento é público, mas a capacidade de processá-lo como insumo para a inteligência tecnológica é, muitas vezes,

<sup>1</sup> <<https://busca.inpi.gov.br/pePI/jsp/programas/ProgramaSearchBasico.jsp>>

restrita por barreiras econômicas.

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. A falta de acesso facilitado a esses dados contribui para a subutilização desse recurso valioso. Esse cenário afeta diretamente os pesquisadores que preferem utilizar ferramentas de acesso aberto, bem como aqueles com recursos financeiros limitados para pesquisa, ou até mesmo carecem de conhecimento técnico para utilizar ferramentas computacionais especializadas, como APIs. Esses perfis de pesquisadores enfrentam desafios consideráveis na condução de seus estudos devido a essas restrições.

Na próxima seção são apresentados os fatores que motivam e de igual modo justificam o desenvolvimento deste estudo.

## 1.2 Justificativa

A temática desta pesquisa surgiu a partir de um problema identificado durante o desenvolvimento da Dissertação de Mestrado do presente pesquisador, defendida em 2022, no qual as bases de dados gratuitas limitavam a recuperação do número total de patentes necessário para a realização do estudo. Ao buscar alternativas, verificou-se que outros pesquisadores enfrentam o mesmo obstáculo em seus trabalhos. A exploração plena do potencial dos dados de patentes é comprometida pela carência de ferramentas e de recursos de acesso aberto adequados para a coleta e análise desses dados, fatores que enfatizam a relevância da pesquisa proposta.

Esta tese insere-se na área de concentração em Sistemas Inteligentes, no âmbito do Programa de Pós-graduação em Modelagem Matemática e Computacional, estabelecendo interface com a Ciência da Informação e a Propriedade Industrial. Trata-se de um estudo de caráter multidisciplinar, voltado para os desafios associados aos repositórios de patentes, em especial no que se refere aos processos de consulta, coleta, armazenamento e análise de dados.

A literatura destaca a crescente importância da análise de patentes em diversas áreas, como inovação, tecnologia e economia. No entanto, também evidencia a subutilização das patentes como uma fonte de pesquisa, indicando que seu potencial informacional ainda é insuficientemente explorado.

A relevância deste estudo reside no fato de propor novas ferramentas, que permitam a coleta abrangente de dados bibliográficos de patentes em repositórios de acesso aberto. Desta forma, esta pesquisa tem o potencial de contribuir significativamente para pesquisas relacionadas à área de propriedade industrial, ao viabilizar o uso eficiente das informações contidas nas patentes e, conseqüentemente, fortalecendo o desenvolvimento tecnológico.

Nesse sentido, a presente tese justifica-se ao formalizar um processo que viabilize coletas mais abrangentes de documentos de patentes em repositórios de acesso aberto, documentando os métodos e recursos utilizados, de modo a estabelecer um protocolo de coleta.

Com base no exposto, na próxima seção são apresentados os objetivos estabelecidos para o desenvolvimento deste estudo.

### 1.3 Objetivos

O objetivo geral desta pesquisa é propor um protocolo sistemático para a coleta, tratamento e persistência de grandes volumes de metadados de patentes oriundos de repositórios de acesso aberto.

Para tanto, será necessário realizar os seguintes objetivos específicos:

- a) Desenvolver um protocolo para a coleta de dados bibliográficos de patentes em repositórios de acesso aberto, de forma a viabilizar a extração de grandes conjuntos de patentes, em conformidade com diretrizes do repositório;
- b) definir critérios objetivos para definir a seleção de repositórios para a coleta de dados bibliográficos de patentes;
- c) construir uma base de dados local;
- d) propor uma estratégia baseada em técnicas de *Large Language Models* (LLM) para viabilizar buscas semânticas, com o intuito de aprimorar a compreensão da linguagem técnica presente nos documentos de patentes.

### 1.4 Público alvo e contribuições esperadas

A presente tese tem como público-alvo interessados em dados provenientes de patentes, abrangendo diversos grupos da sociedade. Entre eles, destacam-se estudantes e pesquisadores, tanto de áreas acadêmicas quanto tecnológicas, que podem utilizar as patentes como fonte primária de informações para suas pesquisas. Além disso, associações empresariais e profissionais da indústria são igualmente contemplados, uma vez que o uso eficaz de informações contidas em patentes pode afetar o desenvolvimento de inovações e a competitividade no mercado. Por fim, a sociedade civil também é considerada, dado que a difusão e o entendimento sobre patentes podem contribuir para o avanço do conhecimento e para a participação mais ativa em debates sobre propriedade industrial e inovação.

No que se refere às contribuições científicas, esta tese pretende:

1. contribuir para a ampliação da eficiência, da abrangência das pesquisas científicas e tecnológicas que utilizam documentos de patentes como fonte de informação, por meio

do desenvolvimento de protocolos e ferramentas que simplificam a coleta, o tratamento e a análise de grandes volumes de dados patentários. Em particular, a proposição de *pipelines* de processamento de dados visa otimizar a extração, a organização e a interpretação das informações tecnológicas contidas nesses documentos;

2. validar um protocolo metodológico para a coleta de grandes volumes de dados bibliográficos de patentes a partir de repositórios de acesso aberto, promovendo a padronização de práticas de extração, validação, integração desses dados. Essa abordagem visa fortalecer a reprodutibilidade, a transparência e a confiabilidade de estudos que utilizam patentes como fonte primária de informação científica e tecnológica;
3. oferecer uma solução computacional automatizada, aberta e reutilizável para a coleta, o armazenamento estruturado e o gerenciamento de metadados de patentes, materializada na ferramenta *Pytente*. Ao reduzir barreiras técnicas associadas ao acesso, à limpeza e à organização dos dados, essa solução amplia o alcance do uso de informações patentárias por públicos diversos, favorecendo a difusão científica e tecnológica em múltiplos contextos institucionais e sociais;
4. mitigar as barreiras impostas pela linguagem técnica e jurídica característica dos documentos de patentes por meio da aplicação de técnicas baseadas em *Large Language Models* (LLMs). Essa abordagem possibilita a recuperação de informações por meio de linguagem natural, tornando os dados patentários mais acessíveis a pesquisadores de diferentes áreas do conhecimento, inclusive àqueles sem formação especializada em propriedade intelectual, e promovendo maior inclusão, interdisciplinaridade e usabilidade dos repositórios de patentes.

## 1.5 Organização do trabalho

O restante desta Tese está organizada da seguinte forma:

- No [Capítulo 2](#), é apresentada uma revisão bibliográfica que aborda os assuntos permeados por esta Tese.
- No [Capítulo 3](#), são apresentados e discutidos alguns trabalhos relacionados aos assuntos abordados nesta Tese.
- No [Capítulo 4](#), são apresentados os procedimentos metodológicos necessários para o desenvolvimento desta Tese.
- No [Capítulo 5](#), são apresentadas todas as etapas do processo de desenvolvimento desta Tese.
- No [Capítulo 6](#), é apresentado o conjunto ferramental projetado para a coleta e tratamento de dados bibliográficos de patentes proposto nesta Tese.
- No [Capítulo 7](#), é apresentado e descrito o conjunto de dados estabelecido, fornecendo um panorama geral de sua estrutura e conteúdo, bem como destacando sua relevância

para os objetivos desta pesquisa.

- Por fim, no [Capítulo 8](#), são apresentadas as considerações finais destacando-se as suas principais contribuições. Esse capítulo apresenta ainda a lista de publicações resultantes dos trabalhos realizados no decorrer da elaboração desta Tese e comenta sobre extensões para possíveis trabalhos futuros.

## 2 Fundamentação Teórica

Este capítulo apresenta um estudo bibliográfico referente aos assuntos abordados nesta Tese. Toda a literatura apresentada tem como objetivo orientar o leitor e o pesquisador. Nas próximas seções, são apresentados: introdução e definição de propriedade intelectual ([Seção 2.1](#)), conceitos referente a documentos de patentes e seus processos ([Seção 2.2](#)), apresentação dos principais repositórios de consulta de documentos de patentes ([Seção 2.3](#)), a exposição de um processo genérico para coleta e análise de documento de patentes ([Seção 2.4](#)), e, por fim, as patentes na perspectiva da Ciência da Informação ([Seção 2.5](#)).

### 2.1 Propriedade Intelectual

A fim de aprimorar a compreensão do conceito de Propriedade Intelectual, é essencial analisar separadamente os termos “propriedade” e “intelectual”. Segundo [Ferreira \(2010\)](#), “propriedade” refere-se ao direito de usar, gozar, dispor de bens e reavê-los de quem os possua injustamente, bem como aos bens sobre os quais se exerce esse direito. Já o adjetivo “intelectual” é definido como relativo à posse de dotes de espírito, de inteligência. Assim, pode-se inferir que Propriedade Intelectual refere-se ao direito sobre tudo aquilo que resulta de inventividade e conhecimento, garantindo ao seu titular o domínio sobre criações intelectuais.

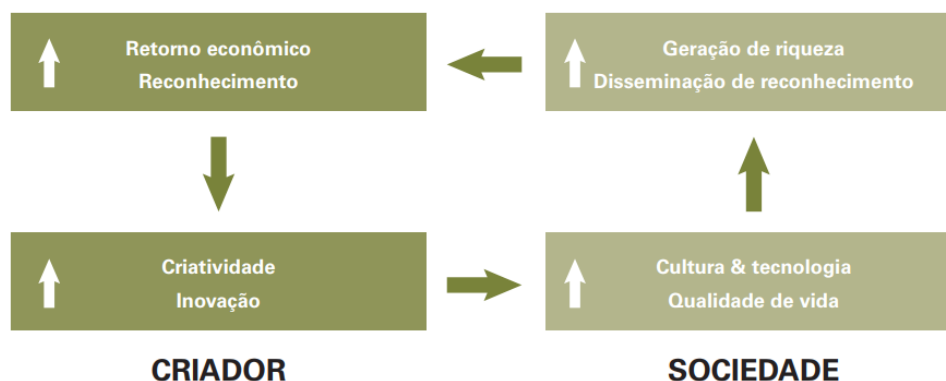
A Organização Mundial da Propriedade Intelectual (OMPI), em inglês *World Intellectual Property Organization* (WIPO), é uma entidade internacional criada em 1967 dentro do sistema das Nações Unidas, incluindo as Uniões de Paris e de Berna, que tem como objetivo a promoção da proteção da propriedade intelectual ao redor do mundo através da cooperação entre Estados ([BARBOSA, 2010](#)). A Convenção da OMPI define como Propriedade Intelectual

a soma dos direitos relativos às obras literárias, artísticas e científicas, às interpretações dos artistas intérpretes e às execuções dos artistas executantes, aos fonogramas e às emissões de radiodifusão, às invenções em todos os domínios da atividade humana, às descobertas científicas, aos desenhos e modelos industriais, às marcas industriais, comerciais e de serviço, bem como às firmas comerciais e denominações comerciais, à proteção contra a concorrência desleal e todos os outros direitos inerentes à atividade intelectual nos domínios industrial, científico, literário e artístico ([BARBOSA, 2010](#)).

Com o objetivo de aprofundar a compreensão sobre as relações entre as diferentes entidades envolvidas no processo de invenção, a [Figura 1](#) apresenta uma representação

gráfica do ciclo que estabelece a interação entre o inventor e a sociedade. Nesse ciclo, o criador contribui para o bem-estar social por meio de uma invenção que promove melhorias na qualidade de vida. Em contrapartida, a sociedade reconhece essa contribuição, gerando um retorno financeiro ao inventor.

Figura 1 – Ciclo de relacionamento entre criador e sociedade

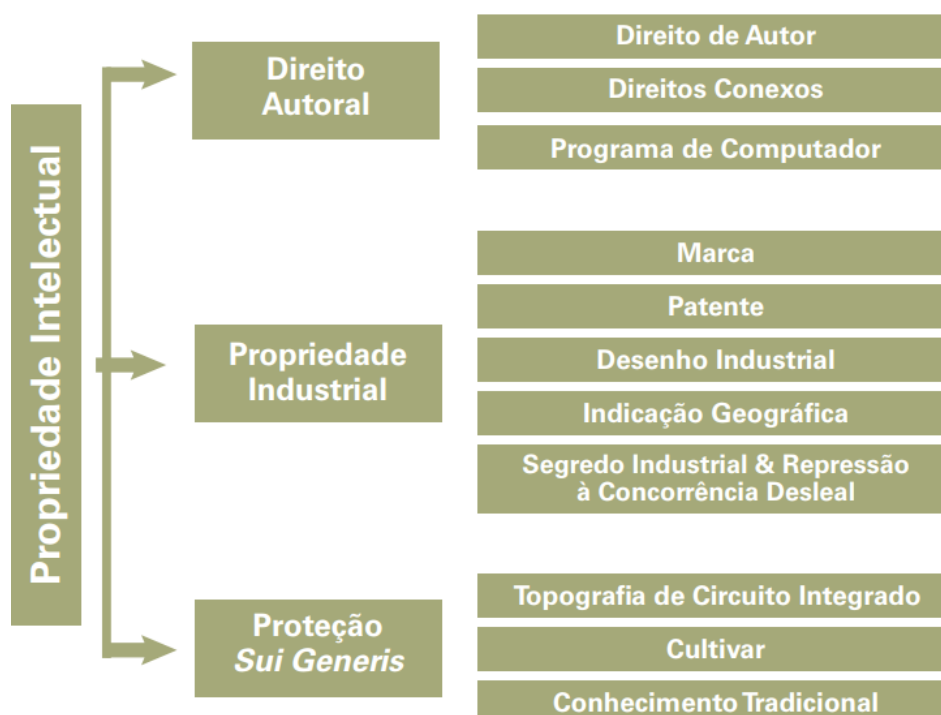


Fonte: [Jungmann e Bonetti \(2010\)](#)

Conforme ilustrado na [Figura 1](#), os inventores e depositantes, identificados como os criadores da patente, desenvolvem produtos e/ou serviços que potencialmente promovem melhorias na qualidade de vida da sociedade. Em contrapartida, esses criadores recebem reconhecimento e retorno financeiro como resultado da inovação depositada. A sociedade, por sua vez, consome esses produtos e/ou serviços, gerando riqueza e contribuindo para o reconhecimento dos inventores. Este ciclo sintetiza, de forma sucinta, o processo de proteção intelectual, conforme definido pela OMPI.

De acordo com [Jungmann e Bonetti \(2010\)](#), a Propriedade Intelectual se divide em três grandes grupos, “direito autoral”, “propriedade industrial” e “proteção *sui generis*”. Neste contexto, [Quoniam, Kniess e Mazieri \(2014\)](#) destacam que patentes, que são parte fundamental deste estudo, pertencem ao grupo “propriedade industrial”, junto com marcas, desenhos industriais e indicações geográficas de origem, conforme é ilustrado na [Figura 2](#).

Figura 2 – Grupos da Propriedade Intelectual



Fonte: Jungmann e Bonetti (2010)

A Propriedade Intelectual é um ramo do Direito que, por meio de um conjunto de leis, assegura aos inventores ou titulares de uma criação o direito de, por um período determinado, obter recompensas por suas inovações. Essas criações podem abranger os campos industrial, científico, literário ou artístico.

No Brasil, diversas leis estabelecem as diretrizes para o direito de propriedade intelectual. Dentre elas, destaca-se a Lei 9.279, de 14 de maio de 1996<sup>1</sup>, que regula direitos e obrigações relativos à propriedade industrial e que substituiu o antigo Código da Propriedade Industrial; Lei 9.456, de 25 de abril de 1997<sup>2</sup>, que é a lei do Cultivares; Lei 9.609, de 19 de fevereiro de 1998<sup>3</sup>, que é a lei de Programa de Computador e, por fim, a Lei 9.610, de 19 de fevereiro de 1998<sup>4</sup>, que altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências, promulgada em substituição a Lei 5.988, de 14 de dezembro de 1973 (BRASIL, 2020a).

Nesse contexto, a Propriedade Intelectual confere aos inventores direitos exclusivos sobre suas criações, o que é fundamental para fomentar a inovação e o desenvolvimento econômico. Esses direitos permitem que os inventores adotem medidas para proteger suas

<sup>1</sup> Disponível em <[http://www.planalto.gov.br/ccivil\\_03/leis/l9279.htm](http://www.planalto.gov.br/ccivil_03/leis/l9279.htm)>

<sup>2</sup> Disponível em <[http://www.planalto.gov.br/ccivil\\_03/leis/l9456.htm](http://www.planalto.gov.br/ccivil_03/leis/l9456.htm)>

<sup>3</sup> Disponível em <[http://www.planalto.gov.br/ccivil\\_03/leis/l9609.htm](http://www.planalto.gov.br/ccivil_03/leis/l9609.htm)>

<sup>4</sup> Disponível em <[http://www.planalto.gov.br/ccivil\\_03/leis/l9610.htm](http://www.planalto.gov.br/ccivil_03/leis/l9610.htm)>

inovações, como proibir terceiros de fabricar, utilizar, comercializar ou importar sua invenção, modelo de utilidade ou desenho industrial sem autorização prévia; impedir a reprodução ou imitação de sua marca; além de adotar ações legais contra a fabricação, importação, exportação, venda, oferta ou armazenamento de produtos com indicação geográfica falsa, entre outras providências. A proteção conferida pela Propriedade Intelectual, além de assegurar uma vantagem competitiva no mercado, também desempenha um papel crucial no incentivo ao investimento em novas tecnologias. A violação desses direitos pode resultar em severas consequências legais, reafirmando a importância de sua proteção.

Tais tópicos são fundamentais, pois estão intimamente relacionados às patentes, que constituem o objeto principal deste trabalho. Embora o presente estudo não enfatize os termos “direitos” e “propriedades”, é importante ressaltar que ambos são essenciais no processo de produção técnica. A Propriedade Intelectual serve como o instrumento legal que assegura a proteção das inovações, garantindo que os responsáveis por essas criações sejam devidamente recompensados.

Na próxima seção, são apresentadas as particularidades de uma patente, tais como a estrutura do documento, o processo de patenteamento e os sistemas de classificação.

## 2.2 Patentes

Conforme apresentado anteriormente, patente é um dos ramos da Propriedade Intelectual. Considera-se patente um documento formal, que é expedido por uma repartição pública, pela qual se conferem e se reconhecem direitos de propriedade e uso exclusivo para uma invenção amplamente descrita (SEBRAE, 2021). O INPI define Patente como

Um título de propriedade temporária sobre uma invenção ou modelo de utilidade, outorgado pelo Estado aos inventores ou autores ou outras pessoas físicas ou jurídicas detentoras de direitos sobre a criação. (BRASIL, 2020b).

As patentes, de acordo com Jurgens e Herrero-Solana (2017), são consideradas documentos abertos ao público que descrevem uma invenção técnica bem detalhada. Em razão de concedida por um escritório de patentes federal ou regional, o proprietário da invenção ganha o direito de explorar comercialmente a invenção por um período de tempo previamente determinado.

Trata-se de um privilégio concedido aos idealizadores, sejam eles pessoas físicas ou jurídicas, que detêm o direito de invenção, abrangendo produtos, processos de fabricação ou aperfeiçoamentos de produtos ou processos já existentes. O inventor oferece à sociedade um novo bem, representado pelo conhecimento, ao divulgar as informações técnicas necessárias para atingir o objetivo proposto. Em contrapartida, a sociedade concede ao inventor

exclusividade na exploração de sua invenção, o que inclui a produção e comercialização do objeto inventado, por um período determinado.

Conforme observado por [Mueller e Perucchi \(2014\)](#), as inovações protegidas por patente possuem potencial valor comercial, o que justifica a existência de uma legislação específica para assegurar sua proteção. O valor de um documento de patente é territorial, ou seja, cada país possui sua própria legislação, bem como um escritório responsável pelo exame e pela concessão de patentes. No Brasil, o Código de Propriedade Intelectual, instituído pela Lei 9.279/96, entrou em vigor em 14 de maio de 1996 e regula os direitos e obrigações relativos à propriedade industrial ([BRASIL, 1996](#)). O Instituto Nacional da Propriedade Industrial (INPI), criado em 1970, é o órgão responsável pela aplicação dessa legislação e atualmente funciona como uma autarquia federal vinculada ao Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC).

Dessa forma, estabelecem-se regras específicas para o pedido de depósito de patente, abrangendo seu processamento e publicação. Entretanto, essas regras variam entre diferentes jurisdições. Na maioria das jurisdições, existe um sistema que prevê a realização de um exame no qual a matéria técnica reivindicada é analisada, verificando se atende a determinadas condições de patenteabilidade, como novidade, atividade inventiva e aplicabilidade industrial. Na jurisdição brasileira, conforme estipulado pela Lei 9.279/96, são estabelecidos os critérios de patenteabilidade, que afirmam que toda invenção que satisfaça os requisitos de novidade, atividade inventiva e aplicação industrial é considerada patenteável.

Além disso, não devem ser classificadas como patenteáveis: descobertas, teorias científicas e métodos matemáticos; concepções puramente abstratas; esquemas, planos, princípios ou métodos comerciais, contábeis, financeiros, educativos, publicitários, de sorteio e de fiscalização; obras literárias, arquitetônicas, artísticas e científicas ou qualquer criação estética; programas de computador em si; apresentação de informações; regras de jogo; técnicas e métodos operatórios ou cirúrgicos, assim como métodos terapêuticos ou diagnósticos para aplicação em seres humanos ou animais; e ainda toda ou parte de seres vivos naturais. É relevante mencionar também que os materiais biológicos encontrados na natureza, ou que dela tenham sido isolados, incluindo o genoma ou germoplasma de qualquer ser vivo natural, bem como os processos biológicos naturais, estão excluídos da patenteabilidade.

De acordo com o [BRASIL \(2020b\)](#), no Brasil, as patentes podem ser divididas em três tipos, a saber:

- **Patente de Invenção (PI):** São os produtos ou processos que atendam aos requisitos de atividade inventiva, novidade e aplicação industrial. Sua validade é de 20 anos a partir da data do depósito.
- **Patente de Modelo de Utilidade (MU):** Objeto de uso prático, ou parte deste, suscetível-

vel de aplicação industrial, que apresente nova forma ou disposição, envolvendo ato inventivo, que resulte em melhoria funcional no seu uso ou em sua fabricação. Sua validade é de 15 anos a partir da data do depósito.

- **Certificado de Adição de Invenção:** Não é bem um tipo, porém um adendo a uma patente existente. Aperfeiçoamento ou desenvolvimento introduzido no objeto da invenção, mesmo que destituído de atividade inventiva, porém ainda dentro do mesmo conceito inventivo. O certificado será acessório à patente e com mesma data final de vigência desta.

Após o vencimento desse período, a patente é caracterizada como domínio público para o livre uso da sociedade.

Para realizar depósito de uma patente, o depositante tem que elaborar o “Pedido de Depósito”, documento este que consiste na elaboração de um relatório, em que o depositante deve descrever de maneira clara e com riqueza de detalhes o objeto ou o processo a ser patentado. Segundo [BRASIL \(2021\)](#), o pedido de patente é composto pelos seguintes documentos:

- **Relatório Descritivo:** É um relatório descritivo elaborado de forma suficientemente clara para que um técnico na área possa reproduzir e indicar a melhor maneira de execução. O relatório destaca o problema identificado no estado da técnica, bem como a solução proposta pela invenção, especificando o setor técnico ao qual se destina. Além disso, demonstra que a solução apresentada constitui uma novidade em relação às outras soluções existentes, ressaltando a evolução tecnológica e as vantagens em comparação com o estado da técnica.
- **Quadro reivindicações:** São apresentadas as particularidades do pedido, apresentando de maneira precisa a matéria objeto da proteção.
- **Resumo:** No resumo, apresenta-se uma descrição sucinta do conteúdo exposto no pedido, com um limite de 50 a 200 palavras, garantindo a clareza sobre o invento. O resumo não deve exceder 20 linhas e é vedada qualquer menção ao mérito ou ao valor da invenção.
- **Desenhos:** Informações complementares, que possam auxiliar na compreensão da invenção, como gráficos, fluxogramas, diagramas, entre outras.

Após a elaboração e submissão do “Pedido de Patente”, caso este seja deferido e o pagamento da retribuição correspondente seja confirmado, é expedida a respectiva carta-patente. De acordo com [Brasil \(1996\)](#), para a concessão de uma patente, é analisado se o pedido atende aos Artigos 10 e 18, que estabelecem as condições sob as quais uma ideia não pode ser patenteada, a saber:

"Art. 10. Não se considera invenção nem modelo de utilidade: I - descobertas, teorias científicas e métodos matemáticos; II - concepções puramente abstratas; III - esquemas, planos, princípios ou métodos comerciais, con-

tábeis, financeiros, educativos, publicitários, de sorteio e de fiscalização; IV - as obras literárias, arquitetônicas, artísticas e científicas ou qualquer criação estética; V - programas de computador em si; VI - apresentação de informações; VII - regras de jogo; VIII - técnicas e métodos operatórios ou cirúrgicos, bem como métodos terapêuticos ou de diagnóstico, para aplicação no corpo humano ou animal; e IX - o todo ou parte de seres vivos naturais e materiais biológicos encontrados na natureza, ou ainda que dela isolados, inclusive o genoma ou germoplasma de qualquer ser vivo natural e os processos biológicos naturais."

"Art. 18. Não são patenteáveis: I - o que for contrário à moral, aos bons costumes e à segurança, à ordem e à saúde públicas; II - as substâncias, matérias, misturas, elementos ou produtos de qualquer espécie, bem como a modificação de suas propriedades físico-químicas e os respectivos processos de obtenção ou modificação, quando resultantes de transformação do núcleo atômico; e III - o todo ou parte dos seres vivos, exceto os microorganismos transgênicos que atendam aos três requisitos de patenteabilidade - novidade, atividade inventiva e aplicação industrial - previstos no art. 8º e que não sejam mera descoberta. Parágrafo único. Para os fins desta Lei, microorganismos transgênicos são organismos, exceto o todo ou parte de plantas ou de animais, que expressem, mediante intervenção humana direta em sua composição genética, uma característica normalmente não alcançável pela espécie em condições naturais."

Após a análise de patenteabilidade, é realizada a avaliação dos documentos conforme as exigências específicas de cada área tecnológica, verificando se a documentação é suficientemente detalhada para que um profissional da área consiga reproduzir os resultados. Uma vez que tudo esteja em conformidade, é concedido um documento oficial denominado carta patente, o qual inclui o número, título e natureza da invenção, o nome do inventor, o prazo de vigência, o relatório descritivo, as reivindicações e os desenhos, além dos dados relativos à propriedade. A concessão é realizada em favor de quem a solicitou, não necessariamente do inventor, formalizando assim a transferência dos direitos sobre a invenção ou modelo.

### 2.2.1 Sistemas de classificação de patentes

As patentes concedidas são classificadas de acordo com sua natureza, conforme o sistema de Classificação Internacional de Patentes (CIP), ou *International Patent Classification* (IPC) em inglês. O CIP é o sistema de classificação adotado no Brasil e em mais de 90 outros países. Segundo Mueller e Perucchi (2014), diversos países se uniram para desenvolver esse sistema, com base no tratado internacional celebrado em 1954, a Convenção Europeia para a Classificação Internacional de Patentes de Invenções. Sob a supervisão da Organização Mundial da Propriedade Intelectual, foi firmado em 1971 o Acordo de Estrasburgo Relativo à Classificação Internacional de Patentes, que entrou em vigor em 1975 (MUELLER; PERUCCHI, 2014). A versão mais recente do Guia Internacional de Patentes, publicada em 2025, está disponível em inglês no site da Organização Mundial da Propriedade Intelectual<sup>5</sup>, e em português no site do Instituto Nacional da Propriedade

<sup>5</sup> <<https://www.wipo.int/publications/en/details.jsp?id=4778&plang=EN>>

Industrial<sup>6</sup>.

De acordo com Brasil (2023), o sistema CIP tem como principais objetivos possibilitar a busca e recuperação de documentos de patentes, organizá-los para facilitar o acesso às informações tecnológicas, auxiliar na investigação do estado da técnica em áreas específicas da tecnologia e servir como base para a elaboração de estatísticas sobre propriedade industrial. Vale ressaltar que uma patente pode receber mais de uma classificação, quantas forem necessárias. Caso não exista uma seção específica que melhor classifique o invento, este será inserido na área mais apropriada.

A CIP é um sistema de classificação hierárquico, que divide o conhecimento tecnológico em oito grandes áreas, denominadas seções. No nível mais alto, existem as seções, representadas pelas letras de A a H; cada seção é dividida em classes, cada classe em subclasses e cada subclasse em grupos, que podem conter subgrupos (WIPO, 2020). As oito seções, primeiro nível hierárquico, são:

- 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.

Assim, as seções podem possuir divisões denominadas subseções, que são definidas por títulos sem a utilização de símbolos de classificação. Por exemplo a Seção “**A** - NECESSIDADES HUMANAS”, que possui quatro Subseções:

- AGRICULTURA;
- PRODUTOS ALIMENTÍCIOS; TABACO;
- ARTIGOS PESSOAIS OU DOMÉSTICOS;
- SAÚDE; SALVAMENTO; RECREAÇÃO.

Dessa forma, cada Subseção, agrupa as classes que compõem a Seção. As classes são o segundo nível hierárquico da CIP e são formadas pelos seus respectivos símbolos, resultado da junção da letra da Seção correspondente à Classe e número inteiro formado por dois dígitos. Por exemplo, as classes da Seção “**A** - NECESSIDADES HUMANAS” que compõem a subseção “PRODUTOS ALIMENTÍCIOS; TABACO” são:

- **A21** - COZEDURA AO FORNO; EQUIPAMENTO PARA PREPARO OU PROCESSAMENTO DE MASSAS; MASSAS PARA COZEDURA AO FORNO;
- **A22** - MATANÇA DE ANIMAIS; BENEFICIAMENTO DA CARNE; PROCESSAMENTO

<sup>6</sup> <<https://ipc.inpi.gov.br/classifications/ipc/ipcpub/?notion=scheme&version=20250101>>

DE AVES DOMÉSTICAS OU PEIXES;

- **A23** - ALIMENTOS OU PRODUTOS ALIMENTÍCIOS; SEU BENEFICIAMENTO, NÃO ABRANGIDO POR OUTRAS CLASSES;
- **A24** - TABACO; CHARUTOS; CIGARROS; DISPOSITIVOS SIMULADORES DE FUMO; ARTIGOS PARA FUMANTES.

Na sequência da hierarquia de classificação estabelecida pelo sistema CIP, o terceiro nível corresponde às Subclasses. A Subclasse é identificada pela combinação de seu símbolo com uma letra maiúscula, oferecendo uma descrição mais precisa de seu conteúdo. Por exemplo, as subclasses que integram a classe “**A21** - COZEDURA AO FORNO; EQUIPAMENTO PARA PREPARO OU PROCESSAMENTO DE MASSAS; MASSAS PARA COZEDURA AO FORNO”, fornecem maior especificidade sobre a área de aplicação, a saber:

- **A21B** - FORNOS DE PANIFICAÇÃO; MÁQUINAS OU EQUIPAMENTOS PARA COZEDURA AO FORNO (equipamentos para cozedura doméstica A47J 37/00; aparelhos de combustão F23; fogões ou fogareiros domésticos sendo total ou parcialmente fornos F24B, F24C);
- **A21C** - MÁQUINAS OU EQUIPAMENTO PARA FAZER OU BENEFICIAR MASSAS; MANIPULAÇÃO DE ARTIGOS COZIDOS FEITOS DE MASSA;
- **A21D** - TRATAMENTO, p. ex. CONSERVAÇÃO, DE FARINHAS OU MASSAS, p. ex. PELA ADIÇÃO DE MATERIAIS; COZIMENTO; PRODUTOS DE PANIFICAÇÃO; CONSERVAÇÃO DOS MESMOS.

O quarto nível da hierarquia no sistema CIP é composto pelos Grupos. A identificação dos Grupos é representada pelo símbolo da subclasse seguido de um número inteiro, que pode conter até três dígitos, e é concatenado com “/00”. O título atribuído ao grupo define de maneira mais específica o seu escopo. Por exemplo, os grupos que integram a subclasse “**A21** - COZEDURA AO FORNO; EQUIPAMENTO PARA PREPARO OU PROCESSAMENTO DE MASSAS; MASSAS PARA COZEDURA AO FORNO”:

- **A21B 1/00** - Fornos de panificação;
- **A21B 2/00** - Aparelhos de cozimento que utilizam aquecimento à alta frequência ou à luz infravermelha;
- “**A21B 3/00** - Peças ou acessórios de fornos;
- “**A21B 5/00** - Aparelhos para cozer artigos especiais; Outros aparelhos para cozimento;
- “**A21B 7/00** - Instalações de panificação.

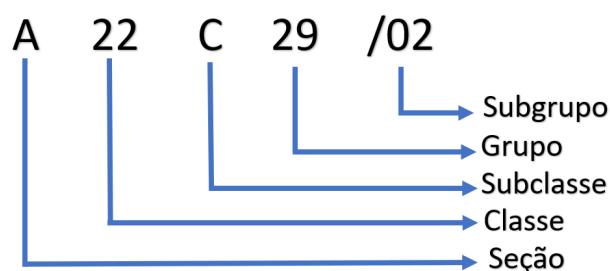
Por fim, a hierarquia do sistema de Classificação Internacional de Patentes (CIP) é concluída pelos Subgrupos. Os Subgrupos são representados pela variação do símbolo de identificação do grupo a qual pertence, ou seja, os valores “00” definidos após a barra são substituídos por um número inteiro que pode ter até três dígitos. Os títulos atribuídos a cada subgrupo definem precisamente um campo de matéria técnica, além de possuir um ou mais

pontos que indicam a posição hierárquica desse subgrupo. Por exemplo, os subgrupos que compõem o grupo "**A21B 1/00** - Fornos de panificação":

- **A21B 1/02** •caracterizados pelas disposições de aquecimento;
- **A21B 1/04** •Fornos aquecidos por fogo somente antes do cozimento;
- **A21B 1/06** •Fornos aquecidos por radiadores;
- **A21B 1/08** ••por radiadores aquecidos a vapor;
- **A21B 1/10** ••por radiadores aquecidos por outros fluidos que não o vapor;
- **A21B 1/14** ••Disposição dos radiadores;
- **A21B 1/22** ••por radiadores elétricos (A21B 2/00 tem prioridade; elementos de aquecimento elétrico H05B);
- **A21B 1/24** •Fornos aquecidos por meios que fluem através dos mesmos;
- **A21B 1/26** ••por ar quente;
- **A21B 1/28** ••por produtos de gases de combustão;
- **A21B 1/33** •Fornos aquecidos diretamente por produtos de combustão (A21B 1/04 tem prioridade);
- **A21B 1/36** •Fornos aquecidos diretamente por fluidos quentes (A21B 1/06, A21B 1/33 têm prioridade);
- **A21B 1/40** •caracterizados pelos meios de regulação da temperatura (elementos sensíveis à temperatura G01K) ;
- **A21B 1/42** •caracterizados por superfícies de cozimento que se movem durante o cozimento (transporte em geral B65G);
- **A21B 1/44** •com as superfícies girando em plano horizontal;
- **A21B 1/46** •com superfícies suspensas por um transportador sem-fim ou uma roda giratória;
- **A21B 1/48** •com as superfícies em forma de cinta sem-fim;
- **A21B 1/50** •caracterizados pelo fato de possuírem superfícies de cozimento removíveis;
- **A21B 1/52** •Fornos portáteis; Fornos desmontáveis (fogões para viagem ou acampamento A47J 33/00).

Diante do exposto, a [Figura 3](#) ilustra um exemplo de classificação atribuída a uma patente, acompanhado de seu respectivo detalhamento.

Figura 3 – Detalhamento de uma classificação de patente



Fonte: Elaboração do autor

- **Seção A:** Necessidades Humanas.
- **Classe 22:** Matança de animais; beneficiamento da carne; processamento de aves domésticas ou peixes.
- **Subclasse C:** Processamento de carnes, aves domésticas ou peixes.
- **Grupo 29:** Processamento de crustáceos, p. ex. ostras, lagostas.
- **Subgrupo 02:** Processamento de camarões, lagostas ou similares.

O sistema CIP não é o único sistema de classificação de patentes. De acordo com [Carvalho e Santos \(2019\)](#), existe o sistema CPC (*Cooperative Patent Classification*, em português, Classificação Cooperativa de Patentes). A CPC é um sistema de classificação criado pelo Escritório Europeu de Patentes (EPO) em parceria com o Escritório Norte Americano de Patentes (USPTO), baseado na CIP, porém com mais detalhes acerca da classificação. Enquanto a CIP possui em torno de 70 mil grupos, a CPC possui em torno de 200 mil grupos. A CPC é adotada pelo INPI, em conjunto com a CIP desde 2014.

As patentes depositadas nos Estados Unidos são classificadas segundo o sistema *United States Patent Classification* (USPC), mantido pelo Escritório Norte-Americano de Patentes. O USPC adota uma classificação própria, que organiza os documentos de patentes em conjuntos relativamente pequenos, baseados em assuntos correlatos. Cada conjunto é composto por um elemento principal denominado “classe” e um elemento secundário denominado “subclasse”. A classe geralmente descreve uma tecnologia específica, enquanto as subclasses detalham processos, características estruturais e funcionais do objeto em questão ([USPTO, 2012](#)).

### 2.2.2 Famílias de patentes

É importante ressaltar que uma patente depositada no Brasil tem validade apenas no território nacional. Para obter proteção em outros países, é necessário realizar o depósito da patente em cada jurisdição desejada, formando assim uma família de patentes. Em termos gerais, uma família de patentes consiste em um conjunto de patentes depositadas em diferentes países com o objetivo comum de proteger a mesma invenção. O direito

de prioridade do primeiro depósito é estendido aos demais países onde os pedidos são realizados. Conforme [INPI \(2020\)](#), uma família de patentes é composta por um conjunto de patentes ou pedidos, nacionais ou internacionais, que compartilham a mesma reivindicação de propriedade ou estão relacionados ao mesmo depósito internacional.

O conceito de família de patentes teve origem na Convenção da União de Paris (CUP), de 1883, que criou um sistema internacional de propriedade industrial com o objetivo de harmonizar os diferentes sistemas jurídicos dos países membros em relação à Propriedade Industrial. Foram estabelecidos quatro princípios básicos, entre eles o princípio da Prioridade Unionista. Esse princípio prevê que, caso sejam realizados depósitos sucessivos de uma mesma invenção em distintos países membros da União, dentro do período máximo de 12 meses a partir da data do primeiro depósito, a invenção estará protegida em todos os países em que o depósito foi efetuado. O conjunto de patentes relacionadas à mesma invenção em diferentes países da União é denominado família de patentes ([WIPO, 2021](#)). Desde o texto inicial de 1883, a Convenção passou por sete revisões: em Bruxelas, em 14 de dezembro de 1900; em Washington, em 2 de junho de 1911; em Haia, em 6 de novembro de 1925; em Londres, em 2 de junho de 1934; em Lisboa, em 31 de outubro de 1958; em Estocolmo, em 14 de julho de 1967; e foi modificada em 2 de outubro de 1979. O Brasil aderiu à União em 1992 ([INPI, 2021a](#)). A Organização Mundial da Propriedade Intelectual disponibilizou em português o documento da Convenção de Paris para a Proteção da Propriedade Industrial, que pode ser acessado no endereço eletrônico <[https://www.wipo.int/edocs/pubdocs/pt/wipo\\_pub\\_201.pdf](https://www.wipo.int/edocs/pubdocs/pt/wipo_pub_201.pdf)> ([WIPO, 2021](#)).

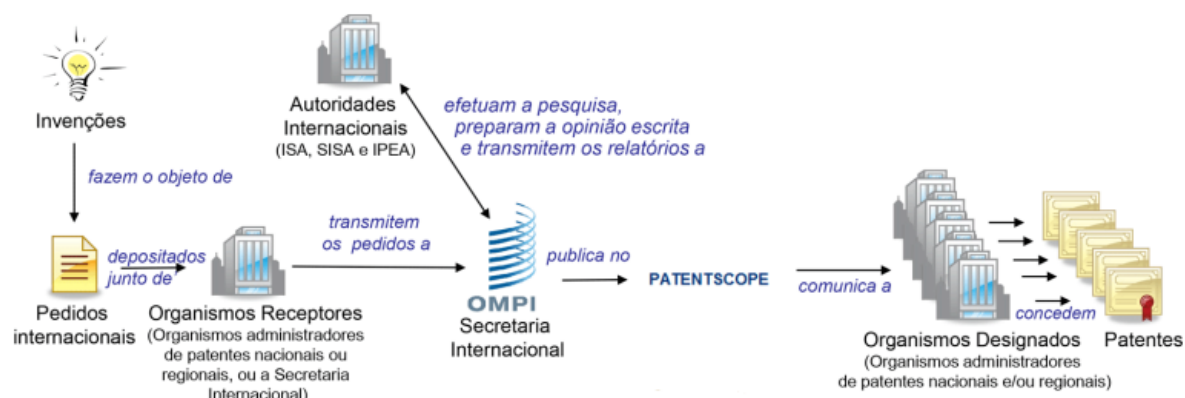
A [WIPO \(2007\)](#) classifica a família de patentes em três grupos distintos:

- **Família de patentes domésticas:** uma família de patentes que consiste exclusivamente em diferentes publicações processuais para o mesmo aplicativo de origem.
- **Família de patente simples:** uma família de patentes relacionada à mesma invenção, composta por membros que compartilham a mesma base de “direito de prioridade”, originada a partir do(s) mesmo(s) pedido(s) inicial(is).
- **Família estendida de patentes:** uma família de patentes relacionada a uma ou mais invenções, em que cada membro tem por base de seu direito de prioridade, pelo menos um pedido originário em comum em pelo menos um outro membro da família.

Também é possível garantir a proteção em mais de um país simultaneamente através do Tratado de Cooperação de Patentes (PCT), administrado pela a Organização Mundial da Propriedade Intelectual. Por meio de um único pedido internacional, é possível garantir a proteção da invenção em mais de 150 países<sup>7</sup> membros do PCT ([WIPO, 2023](#)). A [Figura 4](#) apresenta o resumo do processo.

<sup>7</sup> <[https://www.wipo.int/pct/pt/pct\\_contracting\\_states.html](https://www.wipo.int/pct/pt/pct_contracting_states.html)>

Figura 4 – Resumo do sistema do PCT



Fonte: Elaborado por [WIPO \(2023\)](#)

O PCT simplifica a proteção internacional do invento, removendo a necessidade de apresentar pedidos separados em cada país individualmente. Após concluído o processo de depósito no PCT, todos os escritório de patentes dos países membros do tratado são notificados da nova proteção concedida.

### 2.2.3 Dados bibliográficos da patente

O documento emitido após a conclusão do processo de depósito e concessão de uma patente é composto por informações de natureza jurídica e tecnológica. Do ponto de vista jurídico, trata-se de um direito de propriedade industrial conferido por um período determinado. Sob a ótica tecnológica, o documento descreve de forma abrangente um produto ou processo que, em geral, apresenta uma abordagem inovadora para a realização de uma tarefa ou propõe uma solução técnica original para um problema específico.

De acordo com [Oliveira et al. \(2005\)](#), a estrutura de um documento de patente compreende informações bibliográficas e técnicas. Um documento organizado em seções, incluindo a Folha de Rosto, o Relatório Descritivo, as Reivindicações, os Desenhos e o Resumo. A Folha de Rosto, que corresponde à primeira página do documento, apresenta informações essenciais sobre a patente. Nela, é fornecida uma visão geral organizada, contendo dados como o número da patente, o título da invenção, o nome do inventor, as datas de emissão e de depósito do pedido, o país de origem, as classificações da patente, entre outros, sendo esses dados classificados como bibliográficos. Os dados bibliográficos presentes na Folha de Rosto são identificados por códigos *International Agreed Numbers for the Identification of Data* (INID), ilustrado na [Figura 5](#).

Figura 5 – Pedido de patente US2008/0200653 A1



US 20080200653A1

(19) **United States**  
 (12) **Patent Application Publication**  
**Goddard et al.**

(10) **Pub. No.: US 2008/0200653 A1**  
 (43) **Pub. Date: Aug. 21, 2008**

---

(54) **COMPOSITIONS AND METHODS FOR THE DIAGNOSIS AND TREATMENT OF TUMOR**

(76) Inventors: **Audrey Goddard**, San Francisco, CA (US); **Austin L. Gurney**, Belmont, CA (US); **Paul Polakis**, Burlingame, CA (US); **Victoria Smith**, Burlingame, CA (US); **William I. Wood**, Hillsborough, CA (US); **Thomas D. Wu**, San Francisco, CA (US); **Zemin Zhang**, Foster City, CA (US)

Correspondence Address:  
**HELLER EHRMAN LLP**  
**4350 La Jolla Village Drive, 7th Floor**  
**San Diego, CA 92122 (US)**

(21) Appl. No.: **12/031,612**  
 (22) Filed: **Feb. 14, 2008**

**Related U.S. Application Data**

(63) Continuation of application No. 10/916,250, filed on Aug. 11, 2004, now Pat. No. 7,368,250, which is a continuation of application No. 10/125,166, filed on Apr. 17, 2002, now abandoned, which is a continuation-in-part of application No. 09/918,585, filed on Jul. 30, 2001, now abandoned, which is a continuation-in-

part of application No. 09/904,553, filed on Jul. 13, 2001, now abandoned, which is a continuation-in-part of application No. PCT/US01/06520, filed on Feb. 28, 2001, which is a continuation-in-part of application No. 09/665,350, filed on Sep. 18, 2000, now abandoned, which is a continuation-in-part of application No. PCT/US00/04414, filed on Feb. 22, 2000, which is a continuation-in-part of application No. PCT/US00/04341, filed on Feb. 18, 2000, which is a continuation-in-part of application No. PCT/US99/21090, filed on Sep. 15, 1999, which is a continuation-in-part of application No. PCT/US99/05028, filed on Mar. 8, 1999, which is a continuation-in-part of application No. PCT/US98/19330, filed on Sep. 16, 1998.

(60) Provisional application No. 60/077,649, filed on Mar. 11, 1998, provisional application No. 60/063,542, filed on Oct. 28, 1997.

**Publication Classification**

(51) **Int. Cl.**  
**C07K 16/18** (2006.01)

(52) **U.S. Cl.** ..... **530/387.3; 530/387.9**

(57) **ABSTRACT**  
 The present invention is directed to compositions of matter useful for the diagnosis and treatment of tumor in mammals and to methods of using those compositions of matter for the same.

Fonte: Danilo Zamprônio (2021)

A OMPI é responsável pela definição dos códigos relacionando a Propriedade Intelectual. A Norma ST.9<sup>8</sup> estabelece a especificação dos códigos relacionados a documentos de patentes. O Quadro 1 apresenta uma pequena amostragem dos códigos definidos na norma ST.9.

<sup>8</sup> <<https://www.wipo.int/export/sites/www/standards/en/pdf/03-09-01.pdf>>

Quadro 1 – Códigos INID

| Classe               | INID | Descrição                          |
|----------------------|------|------------------------------------|
| Identificação        | 11   | Número da patente                  |
|                      | 12   | Tipo de documento                  |
| Pedido de Patente    | 21   | Número de depósito                 |
|                      | 22   | Data de depósito                   |
|                      | 25   | Idioma que o pedido foi depositado |
| Publicação           | 45   | Data de publicação                 |
| Informações técnicas | 51   | Classificação CIP e CPC            |
|                      | 54   | Título da invenção                 |
|                      | 57   | Resumo                             |
| Interessados         | 71   | Nome do depositante                |
|                      | 72   | Nome do inventor                   |
|                      | 73   | Nome do titular                    |

Fonte: [WIPO \(2013\)](#).

O Relatório Descritivo, subsequente à Folha de Rosto, apresenta em detalhes todos os procedimentos tecnológicos relacionados à invenção. Essa seção inclui o campo técnico, que oferece uma introdução contextualizando a invenção; os antecedentes da invenção, que constituem uma revisão do estado da técnica disponível na época do depósito do pedido; o sumário da invenção, que fornece uma visão geral resumida, destacando suas características e principais vantagens; uma descrição detalhada e os modos de realização, que trazem exemplos práticos para facilitar a compreensão da invenção ([OLIVEIRA et al., 2005](#)). Cabe ressaltar que a Lei nº 9.279/96 define o estado da técnica como "tudo aquilo tornado acessível ao público antes da data de depósito do pedido de patente, por descrição escrita ou oral, por uso ou qualquer outro meio, no Brasil ou no exterior" ([BRASIL, 1996](#)). Esses elementos constituem as informações técnicas presentes nos documentos de patente.

Na seção de Reivindicações é definida a invenção a ser protegida, considerada uma das partes mais importantes do documento de patente. Com natureza técnico-jurídica, as reivindicações definem a abrangência da matéria reivindicada, ou seja, são estabelecidos os limites da proteção outorgada ao invento. Nas reivindicações, é delineado de forma precisa e detalhada os elementos específicos pelos quais objetiva a proteção legal, podendo incluir componentes estruturais, características funcionais ou métodos específicos associados à invenção. As reivindicações podem ser classificadas em "Reivindicações Independentes", que não dependem de nenhuma outra reivindicação no documento e definem, por si só, o escopo da invenção. Existem ainda as "Reivindicações Dependentes", que dependem de uma ou mais reivindicações anteriores para definir seu escopo e costumam fornecer detalhes adicionais ou limitações específicas à invenção. Assim, as reivindicações devem ser fundamentadas no Relatório Descritivo, delimitando as todas as particularidades do pedido ([OLIVEIRA et al., 2005](#)). Uma patente pode conter uma ou mais reivindicações, entretanto, de acordo com [Lanjouw e Schankerman \(2004\)](#), quanto maior o número de reivindicações maior o valor esperado da patente. Quando apropriado, pode incluir desenhos

técnicos, diagramas ou gráficos para ilustrar melhor a invenção.

E, por fim, o resumo que deve ser iniciado pelo título, em que sua finalidade é oferecer uma compreensão rápida do que a patente abrange, sem a necessidade de examinar todos os detalhes do relatório descritivo e das reivindicações, destacando o invento de proteção, características técnicas, a solução e principais usos. Em suma, o resumo é composto por uma breve descrição da invenção, objetivo da Invenção, principais características, benefícios e aplicações e quando necessário, breve descrição dos desenhos (SAESP/DIRPA/INPI, 2015).

Tendo sido apresentada a estrutura de um documento de patente, a próxima seção discute os repositórios nacionais e internacionais, que oferecem acesso a documentos de patentes.

## 2.3 Repositórios de documentos de patentes

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 (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 (BRASIL, 1996).

Nesse contexto, há diversos repositórios de patentes disponíveis na internet que permitem o acesso a documentos de patentes. Esses repositórios podem ser de acesso aberto ou comercial, 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:

- **EPO**<sup>9</sup> *European Patent Office* (PIRES; RIBEIRO; QUINTELLA, 2020).
- **IPI**<sup>10</sup> *Intellectual Property India* disponibiliza as patentes depositadas na Índia.
- **JPO** (*Japan Patent Office*)<sup>11</sup> 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 (BRANDÃO, 2016).

<sup>9</sup> <<https://www.epo.org/>>

<sup>10</sup> <<https://www.ipindia.gov.in/>>

<sup>11</sup> <<https://www.jpo.go.jp>>

- **KIPO**<sup>12</sup> *Korean Intellectual Property Office* repositórios de patentes da Coreia do Sul ([PIRES; RIBEIRO; QUINTELLA, 2020](#)).
- **pePI**<sup>13</sup> Pesquisa em Propriedade Industrial é o repositório de documentos de patentes do escritório de patentes do Brasil vinculado ao governo federal, com acesso aberto e mantido pelo INPI ([BRANDÃO, 2016](#)).
- **SIPO** (*State Intellectual Property Office of China*)<sup>14</sup> para buscas em patentes depositadas na República Popular da China ([BRANDÃO, 2016](#));
- **USPTO**<sup>15</sup> *United States Patent and Trademark Office* (USPTO) é o repositório de patentes depositadas nos Estados Unidos da América (EUA) ([BRANDÃO, 2016](#)).

Conforme apresentado, através do repositório pePI, disponibilizado pelo INPI, é possível ter acesso a todos os depósitos de patentes realizados no Brasil, desde que não estejam em sigilo. A interface do repositório é totalmente em português, oferecendo duas formas de busca: a busca básica e a busca avançada. A busca básica é recomendada para a pesquisa de pedidos de patentes que já foram depositados e sobre os quais se deseja obter informações sobre seu estado atual ([Figura 6](#)).

Figura 6 – busca básica de patentes no INPI

Fonte: Elaboração do autor

<sup>12</sup> <<https://www.kipo.go.kr/ko/MainApp.do>>

<sup>13</sup> <<https://busca.inpi.gov.br/pePI/>>

<sup>14</sup> <<https://english.cnipa.gov.cn>> ou <<http://english.sipo.gov.cn/>>

<sup>15</sup> <<https://ppubs.uspto.gov/pubwebapp/static/pages/landing.html>>

Conforme apresentado, a busca básica oferece cinco campos de pesquisa. O campo "Contenha o Número do Pedido" permite informar o número do pedido de patente. O campo "Contenha o Nº da Guia de Recolhimento da União - GRU" possibilita inserir o número da Guia de Recolhimento da União, utilizada para requerer o serviço administrativo. O campo "Contenha o Nº do Protocolo" deve ser preenchido com o número atribuído à petição. Por fim, há o campo "Contenha", que permite uma pesquisa mais específica por palavras-chave relacionadas ao pedido de patente. Nesse campo, é possível escolher entre diferentes critérios de busca, como "todas as palavras", "expressão exata", "qualquer uma das palavras" ou "palavra aproximada". A pesquisa pode ser aplicada ao "título", "resumo", "nome do depositante", "nome do inventor" ou "CPF/CNPJ do depositante". Vale destacar que todos os campos preenchidos serão considerados como critérios de união (*AND*) ao realizar a busca.

Já na busca avançada, é possível pesquisar por 14 campos agrupados pelo tipo de informação, "Números", "Datas", "Classificação", "Palavra Chave" e "Depositante / Titular / Inventor" (Figura 7).

Figura 7 – busca avançada de patentes no INPI

BRASIL Acesso à informação Participe Serviços Legislação Canais

Instituto Nacional da Propriedade Industrial  
Ministério da Economia

Consulta à Base de Dados do INPI

[ Início | Ajuda? | Login | Cadastre-se aqui. ]

> Consultar por: **Base Patentes** | Pesquisa Básica | Calendário | Finalizar Sessão

**PESQUISA AVANÇADA**  
Forneça abaixo as chaves de pesquisa desejadas. Evite o uso de frases ou palavras genéricas.

**Números**

(21) Nº do Pedido:  ?

(33)/(31) País/Nº da Prioridade:  ?

(86) Nº do Depósito (PCT):  ?

☐ Calendário de Patentes expiradas/a expirar ?

☐ Patente Concedida ?

+ Datas

+ Classificação

+ Palavra Chave

+ Depositante/Titular/Inventor

Nº de Processos por Página: 20 ▼

pesquisar » limpar

Rua Mayrink Veiga, 9 - Centro - RJ - CEP: 20090-910

Fale conosco

Fonte: Elaboração do autor

No grupo "Números", é possível pesquisar pelo número do pedido de patente, pelo número da prioridade e pelo número do depósito PCT. No grupo "Datas", é viável pesquisar por período de depósito, período de prioridade, período de depósito PCT e período de publicação PCT. O grupo "Classificação" permite a pesquisa pela classificação e por palavras-chave no sistema CIP. No grupo "Palavra-Chave", é possível buscar termos que possam existir no título ou no resumo da patente. Por fim, no grupo "Depositante/Titular/Inventor", é permitido pesquisar pelo nome do depositante, nome do titular, nome do inventor e CPF/CNPJ do depositante.

Após a realização da busca, seja ela rápida ou avançada, é exibida uma página com uma lista de patentes que atendem a todos os critérios especificados. Dessa forma, é possível acessar todas as informações públicas relacionadas a uma patente, como data de publicação, título, resumo e nomes dos inventores, além de todas as petições e documentos vinculados ao pedido.

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, Ribeiro e Quintella \(2020\)](#), as seguintes plataformas: **Espacenet**<sup>16</sup>, **Latipat**<sup>17</sup>, **Patentscope**<sup>18</sup>, **Google Patents**<sup>19</sup>, **Lens**<sup>20</sup>, **Derwent Innovation Index**<sup>21</sup> e **Orbit Intelligence**<sup>22</sup>.

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 ([ESPECENET, 2023](#)). Trata-se de uma das maiores e mais abrangentes bases de dados de patentes, permitindo que os usuários explorem documentos de diferentes países e realizem buscas por palavras-chave, inventores, titulares de patentes, datas e outras categorias relevantes.

Além disso, a plataforma oferece recursos avançados de pesquisa, incluindo a busca por patentes relacionadas a tecnologias específicas. É possível pesquisar em até 12 campos distintos, assim como busca pela classificação de patentes nos sistemas IPC e CPC. A Espacenet disponibiliza os documentos de patentes no idioma original em que foram depositados, permitindo a exportação de até 500 registros nos formatos XLS e CSV ([PIRES; RIBEIRO; QUINTELLA, 2020](#)).

Entretanto, o serviço foi projetado para uso humano, não permitindo consultas auto-

<sup>16</sup> <<https://worldwide.espacenet.com/>>

<sup>17</sup> <<https://lp.espacenet.com/>>

<sup>18</sup> <<https://patentscope.wipo.int/search/en/search.jsf>>

<sup>19</sup> <<https://patents.google.com/>>

<sup>20</sup> <<https://www.lens.org/>>

<sup>21</sup> <<http://apps.webofknowledge.com>>

<sup>22</sup> <<https://www.orbit.com>>

máticas ou recuperação em lotes. Para situações que demandam essa funcionalidade, é recomendado o uso do OPS (*Open Patent Services*), um serviço da web da Espacenet que fornece acesso aos dados armazenados no banco de dados do EPO. Essa possibilidade torna viável o desenvolvimento de aplicativos e robôs de extração automática para o download de conjuntos de dados mais abrangentes, o que, por sua vez, requer conhecimentos em linguagens de programação.

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 em Português em documentos de patentes de 18 países da América Latina e da Espanha. Esta plataforma é resultado do projeto de colaboração Latipat, que teve início em 2003, com o apoio da Organização Mundial da Propriedade Intelectual, do Instituto Europeu de Patentes e da Oficina Española de Patentes y Marcas (OEPM). Ademais, contou com a participação dos serviços de Escritórios de Propriedade Intelectual de países como Argentina, Bolívia, Brasil, Chile, Colômbia, Costa Rica, Cuba, Equador, El Salvador, Guatemala, Honduras, México, Nicarágua, Panamá, Paraguai, Peru, República Dominicana, Uruguai e Venezuela ([LATIPAT, 2017](#)).

O repositório Latipat oferece três interfaces de busca: "Pesquisa Inteligente", "Pesquisa Avançada" e "Pesquisa por Classificação". Na interface de "Pesquisa Inteligente", os usuários podem informar palavras-chave a serem pesquisadas na coleção Latipat de documentos de patente, podendo estas ser inseridas em espanhol ou em português. É importante destacar que as palavras-chave podem ser fornecidas com ou sem acentos gráficos. Para uma pesquisa mais detalhada, utilizando dados bibliográficos específicos, é possível acessar a "Pesquisa Avançada", que permite a pesquisa em 11 campos distintos. Por fim, na interface de "Pesquisa por Classificação", é viável buscar pelos sistemas de classificação CPC e CIP ([LATIPAT, 2017](#)). Na página de resultados, são exibidos apenas os 500 primeiros resultados, e a exportação pode ser feita em lotes de 25 registros nos formatos XLS ou CSV. Além disso, é possível exportar os documentos de patente em formato PDF.

Por sua vez, Patentscope é o repositório de busca de patentes fornecido pela 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. O repositório Patentscope disponibiliza gráficos construídos com base nos resultados da busca realizada com os resultados das buscas. Para tal, o repositório disponibiliza cinco tipos de pesquisa, uma interface para pesquisas simples, uma para pesquisas avançadas, uma para combinação de campos, uma para expansão multilíngue e, por fim, uma para a pesquisa de compostos químicos, em que é necessário autenticar previamente para obter acesso ([ABRANTES et al., 2022](#)).

Na interface de busca simples, é possível pesquisar pelos nomes das pessoas envolvidas, como solicitante, inventor e empresa; pela data de depósito e data de publicação; pelo número de depósito e publicação; pela classificação no sistema CIP; e por campos textuais, como título e resumo. Na pesquisa avançada, pode-se utilizar operadores booleanos e caracteres especiais, o que possibilita a construção de expressões de busca mais específicas. A combinação de campos permite unir os campos de interesse com os conectores “AND” e “OR”.

A pesquisa de expansão multilíngue possibilita a utilização de palavras-chave, com o sistema oferecendo traduções e sinônimos em 13 idiomas distintos, incluindo português, inglês e espanhol. Por fim, na pesquisa por compostos químicos, é possível buscar pela estrutura completa e subestruturas do composto, informando o composto químico, o nome comercial, entre outros (INPI, 2018).

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, atualmente, indexando documentos completos dos escritórios de patentes dos Estados Unidos, Europa, Japão, China, Coreia do Sul, OMPI, Rússia, Alemanha, O Reino Unido, Canadá, França, Espanha, Bélgica, Dinamarca, Finlândia, Luxemburgo, Países Baixos, Áustria, Austrália, Brasil, Suíça e Taiwan. 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, 2023).

No Google Patents, é possível pesquisar por palavras-chave, números de patentes, nomes de inventores e outros critérios para encontrar documentos de patentes relevantes. Além disso, o serviço oferece recursos adicionais, como a visualização de documentos de patentes, análise de citações e informações sobre o status legal das patentes. É possível pesquisar por até 11 campos inclusive pelos os sistemas de classificação CIP, CPC e pelo o sistema Classificação Americana de Patentes – *United States Patent Classification* (USPC) (PIRES; RIBEIRO; QUINTELLA, 2020). Os resultados apresentados, por padrão, são ordenados por relevância, porém é possível ordená-los por data de arquivamento mais recente ou mais antiga. Entretanto, é importante destacar que apenas a patente mais bem classificada de uma mesma família de patentes simples é exibida na lista de resultados. É possível baixar nos formatos CSV, XLSX e SPIF no máximo 1000 resultados (GOOGLE, 2023).

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. Com início no ano 2000 idealizado pela Cambia, uma organização não governamental sediada em Camberra, Austrália, em parceria com a Queensland University of Technology, foi desenvolvido para promover a transparência e a acessibilidade na área de propriedade in-

telectual e pesquisa. O repositório Lens 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. O repositório Leans possui uma base de dados composta por aproximadamente 133 milhões de patentes de mais 104 escritórios. Nele, os dados que compõem a base são disponibilizados pelos os parceiros do projeto, a saber Microsoft Academic<sup>23</sup>, CrossRef<sup>24</sup>, ORCID<sup>25</sup>, PubMed<sup>26</sup>, Impactstory<sup>27</sup>, CORE<sup>28</sup>, European Patent Office (EPO)<sup>29</sup>, United States Patent and Trademark Office (USPTO)<sup>30</sup>, IP Australia<sup>31</sup>, World Intellectual Property Organization (WIPO)<sup>32</sup> (PIRES; RIBEIRO; QUINTELLA, 2020; THE LENS, 2023).

As principais características e funcionalidades do Lens incluem, uma interface de busca simples, interface de busca estruturada, que permite realizar pesquisas mais específicas, uma interface de busca baseada nos sistemas CPC, IPC ou USPC, interface de busca denominada Patcite, que consiste na procura documentos científicos citadas por patentes ou patentes citadas na literatura acadêmica e, por fim, a interface de busca QUT In4M, que apresenta dados das principais organizações de pesquisa que influenciam na indústria e na inovação. Além disso, é possível realizar *download* dos resultados obtidos nos formatos CSV, RIS, BibTeX ou JSON, entretanto, é limitado a 1.000 documentos (PIRES; RIBEIRO; QUINTELLA, 2020).

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. Pesquisa simples, pesquisa de patente citada e pesquisa avançada são as três modalidades de pesquisa disponíveis no repositório. As pesquisas podem ser realizadas em 20 campos distintos, além da possibilidade de pesquisa pelo sistema de classificação CIP e por um sistema de classificação das patentes feita pelos especialistas da Derwent. Os resultados das pesquisas podem ser exportados nos formatos TXT e CSV em lotes de 500 documentos. Ademais, o repositório Derwent Innovation Index disponibiliza ferramentas para análise estatística de patentes pesquisadas (PIRES; RIBEIRO; QUINTELLA, 2020).

Por fim, o repositório Orbit Intelligence, uma solução dedicada a especialistas em

23 <[www.academic.microsoft.com](http://www.academic.microsoft.com)>

24 <[www.crossref.org](http://www.crossref.org)>

25 <[www.orcid.org](http://www.orcid.org)>

26 <[www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed)>

27 <[www.impactstory.org](http://www.impactstory.org)>

28 <[www.core.ac.uk](http://www.core.ac.uk)>

29 <[www.epo.org](http://www.epo.org)>

30 <[www.uspto.gov](http://www.uspto.gov)>

31 <[www.ipaustralia.gov.au](http://www.ipaustralia.gov.au)>

32 <[www.wipo.int](http://www.wipo.int)>

informações de patentes de propriedade da Questel Co, permite pesquisar, monitorar, analisar, categorizar e disseminar informações de patentes (QUESTEL, 2023). Em 5 de janeiro de 2026, o sistema tinha 93 milhões de famílias de patentes e mais de 150 milhões de documentos individuais de patentes fornecidos por 110 escritórios de patentes, disponibilizando dados patentários a partir de 1827 (QUESTEL, 2025). O repositório disponibiliza várias interfaces de busca, dentre elas destacam-se: a pesquisa fácil, pesquisa avançada, pesquisa semântica, pesquisa por requerente e pesquisa por número da patente, nas quais é possível pesquisar em mais de 100 campos distintos. É possível pesquisar pelos sistemas de classificação IPC, CPC, USPC, ECLA. Além disso, fornece informações sobre documentos técnicos não patenteados, que são citados ou referenciados por patentes. O repositório Orbit Intelligence disponibiliza uma ferramenta de análise estatística de patentes que permite a análise de até 2 milhões de documentos e oferece uma variedade de configurações de dados para análise. O repositório permite a exportação de até 20 mil famílias nos formatos TXT, PDF, CSV, XLS, XLSX, XML. Entretanto, não disponibiliza acesso aberto (PIRES; RIBEIRO; QUINTELLA, 2020).

Mediante o exposto, é possível ressaltar que, além de apresentarem recursos semelhantes, cada repositório de patentes tem sua própria interface e algumas funcionalidades exclusivas. A Tabela 1 apresenta um resumo das principais características apresentadas sobre os repositórios internacionais percorridos nesta seção.

Tabela 1 – Resumo das principais características dos repositórios de patentes internacionais

| Repositório              | Acesso | Tamanho da base de dados                           | Limite exportação | Formato exportação            |
|--------------------------|--------|----------------------------------------------------|-------------------|-------------------------------|
| Derwent Innovation Index | Pago   | 35 milhões de famílias e 74 milhões de documentos  | 500 registros     | XLS, CSV                      |
| Espacenet                | Aberto | 149 milhões de documentos                          | 500 registros     | XLS, CSV                      |
| Google Patents           | Aberto | 120 milhões de documentos                          | 1.000 registros   | CSV, XLSX                     |
| Latipat                  | Aberto | Utilizada a base de dados Espacenet                | 500 registros     | XLS, CSV                      |
| Lens                     | Aberto | 133 milhões de documentos                          | 1.000 registros   | CSV, RIS, BibTeX, JSON        |
| Orbit Intelligence       | Pago   | 93 milhões de famílias e 150 milhões de documentos | 20.000 famílias   | TXT, PDF, CSV, XLS, XLSX, XML |
| Patentscope              | Aberto | 115 milhões de documentos                          | —                 | —                             |

Todos os repositórios mencionados estabelecem um limite de *download*, variando em severidade. Alguns, como o *Derwent Innovation Index*, Espacenet e Latipat, apresentam limites mais restritos, enquanto outros, como o *Orbit Intelligence*, são mais flexíveis. No caso do repositório Patentscope, embora não se encontre uma informação explícita sobre o limite

de *downloads* na página do repositório, a seção de "Perguntas Frequentes"<sup>33</sup> destaca que a prática de *downloads* em massa é estritamente proibida, assim como o uso de agentes automatizados. Além disso, é mencionado que dados brutos estão acessíveis mediante compra.

Além dos repositórios já mencionados, há outros, como o "Thomson Innovation", "LexisNexis PatentAdvisor", "Patbase" e o "Intellectual Property Solutions (IPSolutions)". No entanto, apesar da ampla variedade de repositórios de patentes disponíveis, os autores Singh, Chakraborty e Lavina-Vincent (2016) concordam e destacam que nenhum deles abrange a totalidade das patentes publicadas no mundo. Além disso, a precisão dos resultados de busca não é garantida, devido a vários fatores, como o idioma, os critérios de busca oferecidos pelo repositório e o tamanho da base de dados.

Em sequência, na próxima seção, é abordado um processo eficiente para coleta e análise de dados de patentes.

## 2.4 Coleta de dados de patentes

A coleta eficiente de dados de documentos de patentes é essencial para pesquisas que envolvem o campo da Propriedade Intelectual. A definição de bases de dados especializados, o uso de técnicas de mineração de dados e a análise textual, entre outros elementos, compõem um processo multidisciplinar que abrange diversas áreas do conhecimento (RATURI et al., 2010). A complexidade desse processo decorre, em grande medida, da necessidade de lidar com grandes volumes de dados, muitas vezes organizados de maneira heterogênea, e distribuídos em diferentes repositórios nacionais e internacionais. A diversidade de idiomas, as variações nos sistemas de classificação de patentes e os diferentes formatos de apresentação dos documentos tornam a coleta e a análise desses dados desafiadora, exigindo a integração de áreas como o direito de propriedade intelectual, ciência da computação, ciência da informação, estatística e ciência de dados, entre outras.

Ao longo dos anos, diversas ferramentas e técnicas foram desenvolvidas/adaptadas para auxiliar especialistas na análise de documentos de patentes. Entre essas técnicas, destacam-se o uso de técnicas de Processamento de Linguagem Natural (PNL), para a mineração de texto, abordagens baseadas em análise semântica, e métodos visuais de análise de patentes, como a criação de redes de patentes e mapas tecnológicos.

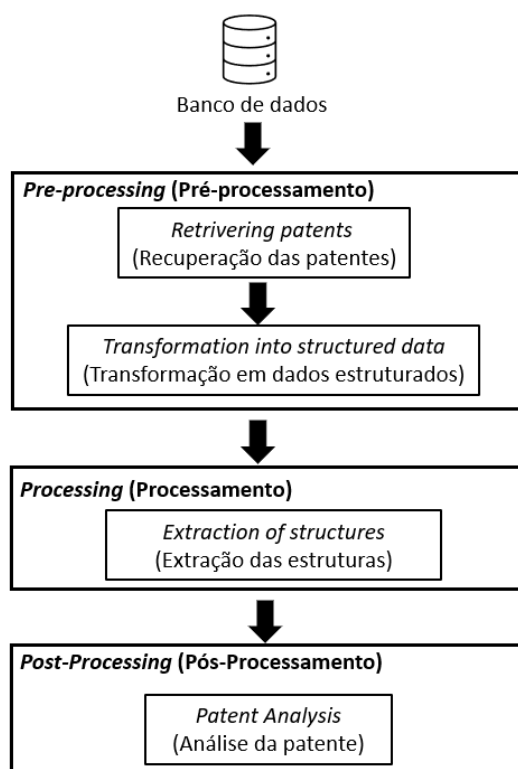
Na Ciência da Informação (CI), especialmente no campo dos Estudos Métricos da Informação, surgiu a Patentometria, uma disciplina dedicada a mensurar o nível de tecnologia e inovação em um país ou setor industrial. Essa avaliação baseia-se na interseção entre o conhecimento científico e sua transição para o domínio tecnológico, realizada

<sup>33</sup> <[https://www.wipo.int/patentscope/en/faqs\\_patentscope.html](https://www.wipo.int/patentscope/en/faqs_patentscope.html)>

principalmente por meio da análise de patentes. No entanto, para que os resultados sejam precisos e relevantes, é imprescindível uma criteriosa seleção dos dados analisados.

Mediante o exposto, é necessário estabelecer um processo sistemático para obtenção dos dados. De acordo com [Moehrle et al. \(2010\)](#), um processo de recuperação de dados de patentes pode ser dividido em três etapas, a etapa de pré-processamento, caracterizada pela coleta dos dados, a etapa de análise dos dados, em que são aplicados métodos e ferramentas de análise e, por fim, a etapa de descoberta do conhecimento envolvendo análise dos resultados de forma visual e mensurável. De igual modo, os autores [Abbas, Zhang e Khan \(2014\)](#) propuseram uma divisão semelhante, uma divisão em camadas classificadas em pré- processamento, processamento e pós-processamento. A [Figura 8](#) apresenta o esboço do processo.

Figura 8 – Processo genérico para coleta e análise patentes



Fonte: Elaboração do autor, adaptado de ([ABBAS; ZHANG; KHAN, 2014](#))

Embora os detalhes específicos possam variar conforme os objetivos da pesquisa e a disponibilidade dos dados, o processo genérico proposto por [Abbas, Zhang e Khan \(2014\)](#) segue etapas bem definidas, começando pela determinação do objetivo e escopo da análise. O objetivo destas é responder perguntas como: "Qual é o objetivo principal da pesquisa?" e "Quais áreas tecnológicas e períodos serão considerados?". Nesse estágio também ocorre a escolha dos repositórios de patentes mais adequados.

Na camada de pré-processamento, são definidos os termos de busca e os métodos

de coleta e armazenamento de dados. Nesta camada, também se realiza a estruturação dos dados, que inclui a eliminação de ruídos, inconsistências e duplicatas, bem como a normalização dos formatos e terminologias utilizados. Em sequência, na camada de processamento, são selecionados conjuntos de dados menores e são extraídos elementos essenciais, como título, inventores, data de publicação e reivindicações. Esta etapa busca organizar os dados de maneira a facilitar a análise posterior.

Por fim, a camada de pós-processamento envolve o exame detalhado e a interpretação dos dados. Isso inclui cálculos estatísticos básicos, como o número de patentes e sua distribuição ao longo do tempo, além da aplicação de técnicas de processamento de linguagem natural (PLN), para a identificação de palavras-chave, frases e temas. Além disso, pode-se realizar o mapeamento das relações entre inventores, empresas e tecnologias, e o agrupamento de patentes com características semelhantes, entre outras análises. Em síntese, a metodologia proposta oferece um roteiro abrangente para explorar o universo das patentes, otimizando a coleta, organização e análise dos dados.

Os processos apresentados por [Moehrle et al. \(2010\)](#) e [Abbas, Zhang e Khan \(2014\)](#) fornecem uma estrutura padronizada que facilita a implementação e a avaliação das técnicas de coleta e análise de dados adotadas. Além disso, esses modelos são flexíveis, permitindo sua adaptação em diversos contextos de pesquisas.

Em uma abordagem computacional, [Silva \(2022\)](#) propôs um processo que permite a recuperação de patentes em repositórios que não oferecem recursos automatizados de coleta. O método sugerido utiliza técnicas de *web scraping* e *web crawler* para reproduzir automaticamente as etapas necessárias à consulta e à visualização dos dados bibliográficos, armazenando em arquivos as patentes de interesse. Assim, para a coleta de patentes no repositório disponibilizado pelo INPI, torna-se necessário automatizar as seguintes etapas:

1. Realizar o login 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” o período de interesse e disparar o evento de pesquisa;
4. Percorrer 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.

A estratégia proposta pode ser adaptada para outros repositórios de patentes,

ajustando-se as etapas necessárias à visualização dos dados bibliográficos. Como demonstrado, a coleta desses dados é uma etapa crucial no processo de análise de documentos de patentes. Portanto, é essencial implementar métodos eficientes e seguros, uma vez que erros na identificação ou interpretação das informações podem resultar em conclusões incorretas e comprometer a validade das análises realizadas.

## 2.5 Análise de patentes com ferramentas da Ciência da Informação

O catálogo das atividades de pesquisa científica de qualquer campo do conhecimento implica em uma seleção criteriosa, diante do grande e crescente volume de informações registradas. Os formatos metodológicos de investigação nos diversos campos do conhecimento se destacam no processo de tratamento, organização, recuperação e transferência das informações. Cada ciência tem suas especificidades em relação à instrumentos de mensuração, por exemplo, a Biometria usada na Biologia, a Econometria usada na Economia, bem como a Cientometria que mede a ciência em um contexto geral (SILVA; BIANCHI, 2001a).

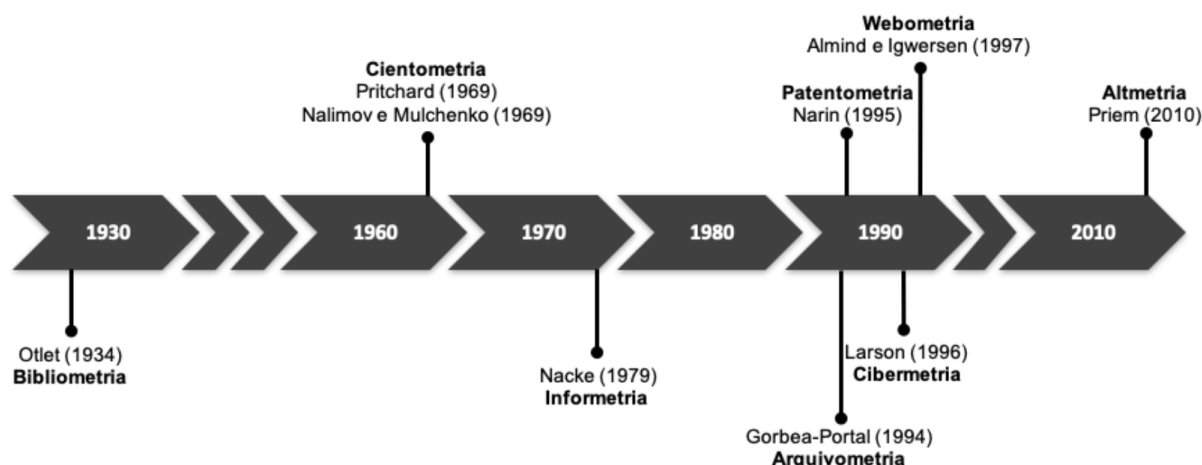
Os Estudos Métricos da Informação são uma das áreas de interesse da Ciência da Informação e têm como foco a identificação e avaliação da informação, seu alcance, influência e impacto. Outras novas técnicas vêm surgindo de maneira proporcional os novos meios de produção e circulação da informação. De 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, Cibernométricos, Webométricos, Patentométricos e Arquivométricos (CURTY; DELBIANCO, 2020).

Cabrini e Gracio (2011) conceituam os estudos métricos como:

Os “Estudos Métricos” compreendem o conjunto de estudos relacionados à avaliação da informação produzida, mais especialmente científica, em diferentes suportes, baseados em recursos quantitativos como ferramentas de análise. Fundamentados na sociologia da ciência, na ciência da informação, matemática, estatística e computação, são estudos de natureza teórico-conceitual, quando contribuem para o avanço do conhecimento da própria temática, propondo novos conceitos e indicadores, bem como reflexões e análises relativas à área. São, também, de natureza metodológica, quando se propõem a dar sustentação aos trabalhos de caráter teórico da área onde estão aplicados.

Conforme destacado por Curty e Delbianco (2020), a bibliometria é o subcampo pioneiro na área dos estudos métricos, sendo seguida pela cientometria. A Figura 9 ilustra uma linha do tempo com os marcos temporais aproximados do surgimento de cada um desses campos de estudo.

Figura 9 – Linha do tempo da evolução dos estudos métricos.



Fonte: [Curty e Delbianco \(2020\)](#)

A bibliometria refere-se ao estudo dos aspectos quantitativos da produção de informação registrada, utilizando padrões e modelos matemáticos para mensurar a produção e disseminação, além de gerar previsões e fornecer suporte para a tomada de decisões ([TAGUE-SUTCLIFFE, 1992](#)). Segundo [Mueller \(2013\)](#), os objetivos da bibliometria incluem a análise e o mapeamento de autorias e coautorias, a caracterização de redes de colaboração, a classificação e descrição da literatura e dos indicadores, a avaliação da produtividade e visibilidade de autores e instituições, bem como a realização de estudos de citação e cocitação. Dessa forma, por meio de análises bibliométricas, é possível, entre outras possibilidades, identificar tendências e áreas do conhecimento emergentes, mensurar a produtividade e reconhecer autores influentes, além de compreender as redes de citação estabelecidas ([GIUSTI et al., 2011](#)).

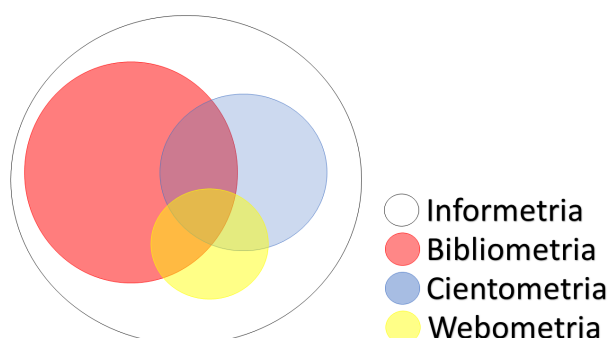
Sob a mesma perspectiva, a cientometria é definida como a análise quantitativa da ciência, voltada para a mensuração do progresso científico e tecnológico. Essa análise baseia-se em fontes secundárias, sem a observação direta do processo de produção de conhecimento e sem a avaliação direta dos resultados obtidos, englobando a bibliometria ([SILVA; BIANCHI, 2001b](#)).

Segundo [Parra, Coutinho e Pessano \(2019 apud VELHO, 1990\)](#), a cientometria busca desenvolver indicadores precisos, por meio de técnicas matemáticas e estatísticas, com o intuito de mensurar e analisar atividades científicas, servindo como parâmetros de avaliação. O termo "cientometria" é a tradução em português de "*scientometrics*" em inglês, onde foi inicialmente mencionado no trabalho de Dobrov e Karennoi, publicado pelo All-Union Institute for Scientific and Technical Information ([VANTI, 2002](#)). Embora a literatura frequentemente utilize o termo "cienciometria", a expressão "cientometria" é uma tradução mais adequada para "*scientometrics*".

A origem do termo "cientometria" remonta à União Soviética e à Europa Oriental, especialmente à Hungria, por volta de 1969, ganhando maior notoriedade em 1977 com o início da publicação da revista *Scientometrics*, originalmente editada na Hungria e atualmente na Holanda (VANTI, 2002).

A cientometria aplica técnicas bibliométricas na ciência, fazendo uso de métodos matemáticos e estatísticos para investigar as características da pesquisa científica. Silva e Bianchi (2001b apud TAUBES, 1993) destacam os principais indicadores bibliométricos utilizados por pesquisadores que estudam a Cientometria, a saber: total de pessoas com titulações acadêmicas ou científicas, o total patentes registradas por cientistas, o total de artigos científicos publicados, o somatório de cientistas que publicam artigos científicos, o total de referências bibliográficas citadas nos artigos científicos, o total de citações recebidas por artigo científico, o total de auxílios à pesquisa recebidos pelos cientistas e a quantidade de recursos destinados às atividades de pesquisa fomentadas pelas agências. Vanti (2002) afirma que a cientometria é composta por outras áreas da Ciência da Informação, como a Informetria, Bibliometria e Webometria, conforme é apresentado na Figura 10.

Figura 10 – Diagrama de inter-relação entre os subcampos



Fonte: Elaborado a partir de (VANTI, 2005)

Ademais, embora o termo "Informetria" seja uma abordagem métrica mais ampla em relação ao escopo, ele foi utilizado pela primeira vez pelo pesquisador alemão Otto Nacke em 1979, em seu trabalho, traduzido para o espanhol, intitulado *"Informetría: un nuevo nombre para una nueva disciplina. Definición, estado de la ciencia y principios de desarrollo"* (CURTY; DELBIANCO, 2020). Neste estudo, o autor apresenta a seguinte definição para o termo: "Informetria é a aplicação de métodos matemáticos aos fatos e situações do campo da informação para descrever e analisar seus fenômenos, descobrir suas leis e servir de apoio para decisões" (NACKE, 1983).

A Informetria refere-se ao estudo dos aspectos quantitativos da informação, independentemente de seu formato. Por essa razão, é considerada a disciplina mais abrangente no campo da Ciência da Informação, pois abrange áreas como bibliometria e cientometria, entre

outras. Além de se restringir a informações científicas, a Informetria também contempla dados relacionados a diversos grupos sociais (TAGUE-SUTCLIFFE, 1992).

Já a Arquivometria, citada pela primeira vez por Gorbea Portal, em 1994, utiliza ferramentas e abordagens estatísticas para otimizar processos relacionados à administração de arquivos. Assim, Portal (2005) define a Arquivometria como a:

aplicação de métodos e modelos matemáticos e estatísticos ao comportamento dos documentos ou manuscritos do arquivo, com o interesse de identificar os fenômenos históricos associados com a estrutura e organização deste tipo de fundo e documentos.

De maneira sucinta, a Arquivometria determina a importância e o valor dos documentos ao longo do tempo, auxiliando na decisão sobre quais documentos devem ser preservados e por quanto tempo. Desenvolve métodos para a organização eficiente de documentos, criando sistemas de classificação que facilitem a recuperação da informação quando necessário. Considerando o avanço da tecnologia, a Arquivometria também lida com questões relacionadas à preservação de documentos digitais, garantindo sua autenticidade e acessibilidade ao longo do tempo.

Em sequência, na linha do tempo elaborada pelos os autores Curty e Delbianco (2020), nos anos 90, surge o emprego do termo "Patentometria". Diversos trabalhos têm utilizado o termo Patentometria para descrever os estudos métricos de documentos de patentes. A patentometria faz parte do grupo que constitui a Bibliometria, pois para a análise de documentos de patentes, são utilizadas adaptações dos indicadores utilizados na Bibliometria (FAGUNDES et al., 2014). Nascimento e Speziali (2020) definem patentometria como a ciência de analisar informações contidas em documentos de patente buscando identificar possíveis relacionamentos e tendências, por meio de estratégias, tais como:

- i. **Inteligência de patentes:** "Uso de informações de patentes para identificar os recursos técnicos de uma ICT (Instituições de Ciência e Tecnologia) ou NIT (Núcleo de Inovação Tecnológica) e o uso dessa inteligência para desenvolver um planejamento estratégico tecnológico";
- ii. **Mapeamento de patentes:** "Usa dados de patentes para criar representações gráficas que mostram um panorama tecnológico de uma área de conhecimento ou de uma invenção";
- iii. **Análise de citações de patentes:** "Auxilia na valoração de uma tecnologia e na identificação de parceiros e no licenciamento novas descobertas tecnológicas".

O conceito de patentometria foi introduzido em 1995 por Narin (NARIN, 1995). Entretanto, o termo foi formalmente estabelecido por Sánchez (1999), que define patentometria como uma abordagem que integra conhecimento certificado e participação em inovação industrial (CURTY; DELBIANCO, 2020). Por sua vez, Aristodemou e Tietze (2018) definem

a patentometria como a ciência que analisa grandes volumes de dados sobre propriedade intelectual, com o objetivo de descobrir relacionamentos, tendências e padrões. Esta análise utiliza matemática, estatísticas, programação computacional e pesquisa operacional, visando fornecer informações que apoiem a tomada de decisões. Ademais, a patentometria é considerada uma abordagem inovadora na ciência da informação.

Todo o conhecimento gerado pela patentometria pode ser aplicado a diversas organizações que se dedicam à avaliação de tecnologia e seu impacto na sociedade. Governos e agências, assim como empresas privadas, podem obter uma perspectiva valiosa sobre determinados tipos de tecnologia ou desenvolvimento (TRIPPE, 2015). Para a construção dos indicadores, a patentometria utiliza diversas fontes de informação, incluindo os dados contidos nos códigos INID presentes nos documentos de patentes (SILVA; NASCIMENTO, 2020). A **Quadro 2** apresenta sete desses indicadores, derivados dos códigos INID, os quais são considerados "indicadores primários", por estarem diretamente explícitos em um documento de patente.

Quadro 2 – Amostragem de indicadores derivados dos código INDI

| Descrição do indicador                                                                                                                                                                                                                    | Campo da patente                                                     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| Avaliação global dos depósitos de famílias de patentes ao longo dos anos, também é conhecida como série temporal.                                                                                                                         | (32) Data de depósito da primeira solicitação (data de prioridade)   |
| Avaliação dos mercados de proteção potenciais, realizada por meio da análise dos países onde os depósitos de patentes foram feitos                                                                                                        | Famílias de patente                                                  |
| Avaliação dos depositantes das patentes (uma atenção especial a existência de diferentes grafias de um mesmo depositante).                                                                                                                | (71) Nome do depositante                                             |
| Avaliação das áreas técnicas os quais os depósitos estão associados (baseados nos sistemas de classificação IPC e CPC)                                                                                                                    | (51) Classificação Internacional                                     |
| Avaliação dos centros produtores de tecnologias. São utilizados os dados do escritório de onde ocorreu o depósito mais antigo. As duas primeiras letras do campo INID-31 (Número designado ao primeiro depósito (prioridade do documento) | (31) Número designado ao primeiro depósito (prioridade do documento) |
| Análise da atratividade de mercado. A razão entre total de depositantes de residentes por total de depósitos de estrangeiros em um dado país.                                                                                             | (71) Nome do depositante                                             |
| Análise de nuvem de palavras, baseado nos campos do título, resumo ou reivindicações.                                                                                                                                                     | (54) Título da invenção / (57) Resumo do conteúdo do documento       |

Fonte: adaptado de Silva e Nascimento (2020).

Os indicadores gerados pelos campos INID revelam, de maneira direta, diversos fatos e tendências, uma vez que estão explicitados nos documentos de patente. Os campos INID contêm dados precisos e objetivos sobre a patente, como a data de depósito, o inventor, o título e a classificação tecnológica, entre outros. Essas informações são cruciais para compreender o contexto da invenção e suas características. Ademais, a organização dos campos INID segue uma estrutura lógica e bem definida, facilitando a extração e a análise dos dados. Além disso, muitos dos dados presentes nos campos INID são quantitativos, como o número de reivindicações e o número de inventores, o que possibilita a realização de análises estatísticas.

O INPI disponibiliza periodicamente diversos indicadores de propriedade industrial que fornecem uma visão estatística do uso da propriedade industrial no Brasil. O [Quadro 3](#) apresenta a descrição de sete dos indicadores disponibilizados pelo o INPI.

Quadro 3 – Indicadores patentométricos INPI

| Indicador                            | Descrição do indicador                                                                               |
|--------------------------------------|------------------------------------------------------------------------------------------------------|
| Total de depósitos de patentes       | Calculado usando como referência a data de protocolo do pedido no INPI                               |
| Ranking de Depositantes residentes   | Usando como critério o primeiro depositante informado no documento de depósito residente no país     |
| Ranking de Depositantes estrangeiros | Usando como critério o primeiro depositante informado no documento de depósito não residente no país |
| Campo tecnológico                    | A primeira classificação CIP recebida pela a patente                                                 |
| Total de depósitos por Município     | Cidade do primeiro depositante da patente                                                            |
| Total de depósito por tipo de pessoa | Tipo de pessoa que realizou o depósito, Pessoa Física, Microempresa assim definida, Pessoa Jurídica  |
| Séries Temporais                     | Depósitos, concessão, publicação de patentes ao longo dos anos em um intervalo de tempo              |

Fonte: adaptado de [INPI \(2021b\)](#), [INPI \(2023\)](#).

O quadro apresentado lista sete indicadores patentométricos utilizados pelo INPI, acompanhados de suas respectivas descrições. Essas descrições exemplificativas oferecem uma visão sucinta do objetivo de cada indicador.

Em suma, a patentometria fornece uma base sólida para a geração de inteligência de patentes, pois possibilita a transformação de dados brutos em informações significativas e acionáveis. Ao integrar dados estruturados e ferramentas analíticas, a patentometria aplica métodos sistemáticos de análise quantitativa e qualitativa, permitindo uma compreensão aprofundada dos dados de patentes. Essa abordagem viabiliza a obtenção de informações valiosas sobre a atividade inventiva, as tendências tecnológicas e o panorama competitivo de um determinado setor.

## 3 Trabalhos Relacionados

Este capítulo apresenta a síntese de um conjunto de trabalhos relacionados aos temas abordados neste Tese. Nas seções subsequentes, são discutidos os principais estudos que tratam das patentes como fonte de dados para análise ([Seção 3.1](#)) e os diferenciais propostos por este trabalho ([Seção 3.2](#)).

### 3.1 Patentes como objeto de pesquisa

Estudos que utilizam informações obtidas por meio de documentos de patentes adotam diversas abordagens, sejam de caráter quantitativo, qualitativo, tecnológico, voltado à gestão da inovação, social, entre outros. Dada a amplitude do tema, os termos de busca utilizados foram: (Patent\* and (database or repository or “banco de dados” or repositório)). Os critérios para a seleção dos estudos incluíram: publicações disponíveis na web em português ou inglês; publicadas nos últimos 20 anos (para considerar trabalhos mais recentes); e que analisassem documentos de patentes de qualquer natureza. A pesquisa foi conduzida entre 2022 e 2024, utilizando os seguintes repositórios: Arca (Fiocruz), Biblioteca Digital de Teses e Dissertações, Google Acadêmico, Microsoft Academic, Portal de Periódicos CAPES/MEC e SciELO.

O trabalho de [França \(2007\)](#) apresenta as patentes como fonte de informação tecnológica, em que é apresentado o contexto histórico da propriedade industrial, abordando conceitos importantes como marcas e patentes, dando maior ênfase neste último. Apresenta um panorama geral sobre o documento de patente e suas particularidades, como o sistema de classificação e objeto que podem ser patenteados. Destaca, ainda, a importância de considerar os dados contidos em documentos de patentes como fonte de informação primária a respeito do desenvolvimento tecnológico. Em seu trabalho, o autor aponta a patente como uma rica fonte de informação tecnológica.

[Quoniam, Kniess e Mazieri \(2014\)](#) abordam em seu trabalho as dimensões da propriedade intelectual, conceituando e destacando patente como rica fonte de dados para pesquisas no campo das Ciências Humanas e Sociais, particularmente em Ciência da Informação e Comunicação, bem como a importância de utilizá-las na transferência de tecnologia entre países e organizações, uma vez que as informações contidas em documentos de patentes, em raros casos, são reaplicadas em outras fontes de informação. Em seu trabalho, os autores não realizam de fato uma análise específica a respeito dos dados de patente, entretanto, enfatizam a importância em utilizar as informações contidas em documentos de patentes e apresentam diversas estratégias de análise dos dados, apontando direcionamentos para pesquisadores interessados no assunto.

Ribeiro e Garcia (2014) apresentam em seu trabalho um panorama da patentometria em periódicos internacionais da ciência da informação. Iniciam seu trabalho citando a bibliometria e a cientometria como as metodologias mais relevantes nos estudos da informetria, por abranger grande área de estudo. Complementando, destacam que a patentometria como uma ferramenta de inteligência competitiva em ciência e inovação, fornece indicadores como base substancial para a tomada de decisão, uma vez que a patentometria é caracterizada por estabelecer indicadores que buscam identificar atividades de inovação e tecnologias, através das informações extraídas de documentos de patentes. Para estabelecer um panorama geral, os autores realizaram um estudo no âmbito internacional, verificando o estado da arte das pesquisas em patentometria na área da ciência da informação. O estudo foi realizado por meio de pesquisas em periódicos com conceitos Qualis A1 a B2 compreendido entre os anos de 1998 a 2013, por termos referente a patentometria, tais como: *patent*, *industrial property*, *innovation patent* e *analysis patent*. Ao concluírem o trabalho, os autores afirmam que a análise de patentes constitui uma ferramenta importante de competitividade para pesquisas, que permite identificar e converter o conhecimento científico em conhecimento tecnológico.

Já Cattivelli (2016) apresenta, em sua dissertação de mestrado, uma análise das patentes concedidas, das Universidades Públicas Brasileiras com o objetivo de identificar fatores e/ou características dessas que incentivam o depósito e, por consequência, a concessão de patentes. A coleta dos dados de patentes para análise foi realizada na base de dados do INPI, na qual foi possível coletar dados de patentes de 29 universidades, do total de 106 universidades pesquisadas, 65,2% corresponde a universidades da região sudeste do país. Com base na CIP (Classificação Internacional de Patentes) das patentes coletadas, foi possível observar que a área do conhecimento com maior quantidade de patentes concedidas são “Química e Metalurgia”, em que os pesquisadores que mais patenteiam, foram professores com formação em Química ou Engenharia Mecânica. A autora ainda destaca que existe uma forte presença do setor público, pois, em sua maior parte, a cotitularidade é constituída por instituições de fomento e de pesquisa. Concluindo, os autores destacam a importância dos Núcleos de Inovação Tecnológica (NITs) nas instituições públicas como atores que contribuem para implantar a cultura do patenteamento no país que, por consequência, contribuem de forma indispensável com a evolução tecnológica do país.

A ferramenta Patent2Net (P2N) é descrita no trabalho “*A new patent processing suite for academic and research purposes*” dos autores Reymond e Quoniam (2016). Consiste em conjunto ferramental de dado patentométricos, com o objetivo de contornar a deficiência existente no processo de ensino e pesquisas acadêmicas sobre Propriedade Intelectual. O P2N é uma ferramenta de código aberto, escrita na linguagem de programação Python, desenvolvido em 2014 no âmbito do programa de mestrado em Economia e Inteligência Territorial da Universidade de Toulon na França. Disponível para os sistemas operacionais

Linux e Windows, o P2N atua como um minerador de documentos de patentes que estão disponíveis na base de dados da Espacenet. Assim, viabiliza acesso a um grande volume de informações patentárias, de forma rápida e organizada, oportunizando o processo de análise dos documentos de patentes.

Mendes e Melo (2017) propõem em seu trabalho uma metodologia para avaliar a tecnologia, por meio de técnicas de cientometria e patentometria, servindo assim como ferramenta de apoio a tomada de decisão para as organizações no tangente a introdução de novas tecnologias. O modelo é composto por 9 etapas, (1) Definição de termos, em que serão definidas as palavras ou expressões, sobre as quais se deseja obter maiores informações ; (2) Validação dos termos por especialistas, em que os termos definidos são revisados e aprovados por especialistas na área afim; (3) Definição das bases a serem utilizadas, por meio de critérios previamente estabelecidos; (4) busca pelos termos nas bases definidas; (5) extração dos registros encontrados, para formar uma base de dados local, para viabilizar as análises; (6) realização da limpeza do dados, correção de possíveis erros de digitação, duplicidades e padronizações; (7) obtenção de listas, gráficos e tabelas; (8) Análise das informações geradas; (9) Conclusões e resultados. Os autores ainda argumentam que a proposta permite fornecer aos tomadores de decisão bases de dados científicas que servirão como fundamentação para tomada de decisões, como indícios de ofertas de conhecimento para o desenvolvimento da tecnologia, bem como sua demanda.

Os autores Scartassini et al. (2018) apresentam uma análise patentométrica sobre as patentes brasileiras com extensão de depósito via Tratado de Cooperação de Patentes (PCT). Um trabalho patentométrico quantitativo de natureza descritiva. Os autores coletaram os dados em março de 2017, dividindo a metodologia em três etapas distintas, a primeira recuperando todas as patentes depositadas no Brasil indexadas na base Derwent Innovation Index (DII) entre o período de 2004 a 2016; na segunda etapa, dentro do conjunto coletado, foram extraídas as patentes com registro prioritário no Brasil; e, por fim, na terceira etapa foi extraído desse conjunto as patentes com depósito prioritário no Brasil e que realizaram pedido de proteção internacional na via PCT. Totalizando 190 patentes consideradas em seu estudo. No processo de análise, foram levantados os maiores depositantes, análise temporal dos depósitos de patentes, o quantitativo das classificações recebidas pelas patentes, concluindo que o depósito via PCT ainda não é um prática comum para os depositantes brasileiros.

Menezes (2019), em sua dissertação de mestrado ao Programa de Pós-graduação Profissional em Gestão, Pesquisa e Desenvolvimento na Indústria Farmacêutica, do Instituto de Tecnologia em Fármacos na Fundação Oswaldo Cruz, realizou um estudo patentométrico em que e almejou-se mapear o cenário tecnológico do desenvolvimento de novas alternativas terapêuticas para o tratamento da tuberculose. Os dados para o estudo foram obtidos por meio de consultas no repositório Derwent World Patents Index, através do Portal

de Periódicos da Capes em setembro de 2018, dentro do período de 2008 a 2018, estabelecendo como critério de seleção as patentes cujo título ou resumo apresentem termos relacionados à tuberculose e que possuam as classificações A61K-039/04 ou A61P-031/06, obtendo um conjunto de 517 patentes consideradas no estudo. Mediante a análise dos indicadores quantitativos, o autor destaca um declínio no número de depósitos por ano, entretanto, foi possível observar a presença tanto de instituições públicas quanto privadas participando ativamente nos processos de depósito de patentes e estudos clínicos.

A pesquisa realizada pelo os pesquisadores [Moura et al. \(2019\)](#) tem como principal objeto a analisar diferentes bases de dados depositantes de patentes tendo como objetivo principal a análise das características inerentes ao processo de coleta e recuperação de patentes disponíveis nos repositórios Derwent Innovations Index, Orbit, INPI, Google Patents e PatentScope. Mediante a buscas pelas patentes depositadas pela Universidade Federal do Rio Grande do Sul (UFRGS), o estudo tem como objetivo explorar as diferenças existentes entre as bases de dados mencionadas. Assim, os pesquisadores puderam constatar que as bases de dados analisadas apresentaram características distintas umas das outras. A DII e a Orbit apresentaram os maiores números de registros recuperados, entretanto, foi possível observar que a DII apresentou algumas inconsistências nos dados retornados. O INPI e o PatentScope apresentaram números de registros mais condizentes com a realizada pela Universidade. O Google Patents, se apresentou com uma fonte não recomendada, dado o tipo de pesquisa realizada, uma vez que não indexa o escritório brasileiro. E, por fim, os autores enfatizam que é necessário examinar com cautela a fonte de dados antes de realizar qualquer tipo de estudo envolvendo documentos de patentes.

Já os autores [Junior e Kühl \(2020\)](#), realizaram um estudo bibliométrico, por meio de pesquisa nos repositórios da SciELO e Capes no período de 2005 a 2019, complementado por um estudo patentométrico por meio da busca na base de patentes do Instituto Nacional da Propriedade Industrial (INPI) no período de 2005 a 2018, tendo como principal objetivo mapear as inovações tecnológicas aplicáveis nas Ciências Contábeis. Na etapa metodológica, foram estabelecidos os critérios de busca e seleção dos dados nos repositórios da SciELO e Capes e INPI, considerando os termos “Contabilidade”, “Contabilidade *and* Inovação”, “Contabilidade *and* Tecnologia” e “Contabilidade *and* Tecnologia da Informação”, com caráter quantitativo. É importante destacar que, através das patentes identificadas, por meio da busca realizada, foram obtidos 10 resultados, dos quais todas as 10 patentes referem-se a Sistemas de Informação aplicáveis à contabilidade, tais como controle financeiro, controle de despesas e gastos, emissão de notas fiscais e o ensino de contabilidade. Os autores concluem considerando que os resultados do estudo bibliográfico estão em concordância com os resultados bibliométrico e patentométrico obtidos, destacando que, em maior parte, os dados levantados nos repositórios considerados no estudo, tratam-se de sistemas informatizados, associado a era da informatização e dos processos computadorizados nas Ciências Contábeis

Em [Silva e Nascimento \(2020\)](#), os autores apontam a patentometria como ferramenta indispensável para a proposição de tecnologias para indústria química, pois a prospecção, baseada em métodos quantitativos e qualitativos, é, de fato, importante para viabilizar a conceitualização de propostas plausíveis no estado da arte. Em sequência, exploram alguns indicadores apresentando o processo de construção de sua relevância no panorama tecnológico. Os autores concluem, mencionando que a patentometria é uma área pouco explorada no Brasil, apesar de seu valor competitivo.

Os autores [Souza et al. \(2021\)](#) abordam o mapeamento do depósito de patentes, marcas, desenhos industriais e registros de programas de computador registrados no INPI por discentes do Programa de Pós-Graduação em Ciência da Propriedade Intelectual da Universidade Federal de Sergipe, com o objetivo de avaliar sua representatividade. Os dados foram obtidos no repositório do INPI, no período de 2014 a 2020. Mediante a análise dos indicadores quantitativos levantados, os autores evidenciaram a existência de registro de propriedade intelectual em todas as categorias pesquisadas caracterizando assim a representatividade dos discentes.

Os pesquisadores [Cunha, Volpato e Pedron \(2023\)](#) propuseram, em seu estudo, o desenvolvimento do “Protocolo de Revisão Sistemática de Literatura Patentária” ou, simplesmente, “RSLP”, com objetivo de amparar com dados para pesquisas na área das Ciências Sociais Aplicadas. Neste estudo, foi realizado o levantamento de estratégias de buscas em bases patentárias, para assim compreender suas particularidades. Fundamentado no método de pesquisa *Design Science Research* (DSR) os autores propuseram, partindo de um conjunto de problemas práticos e teóricos identificados, o método sistemático RSLP, que busca mitigar a complexidade de recuperação de dados de patentes em base de dados como a Espacenet e INPI. Por intermédio de oficinas realizadas com alunos de pós-graduação e pesquisadores, o método proposto foi validado. Como resultado, foi produzido um programa de computador de acesso via internet, que permite aos usuários elaborarem estratégias de consultas de patentes. Contudo, os autores enfatizam que as estratégias elaboradas pelo o protocolo RSLP se aplicam somente às bases de dados do INPI e Espacenet.

No trabalho realizado, pelos autores [Rezende et al. \(2023\)](#), intitulado “Elaboração de um Conjunto de Dados sobre o Registro de Patentes no Brasil”, é apresentado um processo sistemático de construção de uma base de dados de patentes depositadas no escritório de patentes, o INPI. Os dados, que compõem a base de dados apresentada, são provenientes da Revista de Propriedade Industrial, que é disponibilizada semanalmente pelo INPI. Partindo da edição 2474 publicada em junho de 2018, por ser a primeira versão a disponibilizar os dados da revista no formato XML. Enfim, a base de dados proposta agrega 408.705 processos de depósito de patentes, 113.628 titulares, sendo 47.390 brasileiros e 411.724 inventores.

Neste mesmo contexto, no trabalho “Desenvolvimento de um Sistema para Gestão de Propriedade Intelectual para Instituições de Ciência e Tecnologia”, dos autores [Teles et al. \(2023\)](#), é apresentada a proposta de um sistema para gestão dados de patentes denominada “Sistema Integrado de Gestão de Propriedade Intelectual (SIGPPI)”, desenvolvida com a linguagem de programação Python com auxílio do *framework* Django e o sistema de gerenciamento de banco de dados MariaDB. A proposta tem, como principal objetivo, objetivo auxiliar as Instituições de Ciência e Tecnologia no processo de análise e busca de tecnologias. Entre as funcionalidades disponíveis no SIGPPI está a coleta de dados no formato XML, provenientes da Revista de Propriedade Industrial disponibilizada pelo INPI, para a construção da base de dados do sistema.

## 3.2 Considerações sobre os trabalhos relacionados e diferenciais propostos

Na literatura revisada, observa-se um esforço significativo, por parte dos pesquisadores, em destacar os documentos de patentes como uma fonte valiosa de informações tecnológicas. Esses estudos apresentam suas perspectivas com base em conjuntos específicos de patentes, como áreas de pesquisa ou setores tecnológicos, além de proporem alternativas para fomentar o uso de patentes em pesquisas acadêmicas. No entanto, há um ponto comum entre os trabalhos: o foco em subconjuntos, ou seja, no uso de pequenos grupos de dados de patentes, se comparado ao volume total de patentes disponíveis nacionalmente. Diferentemente das abordagens apresentadas, este estudo contempla uma análise detalhada dos processos envolvidos na escolha de repositórios de patentes, na coleta de dados bibliográficos e no armazenamento de grandes conjuntos de dados bibliográficos de patentes.

Distinto dos trabalhos apresentados, este estudo estabelece um conjunto de estratégias para a recuperação e armazenamento de dados bibliográficos de patentes e de dados complementares, culminando em um protocolo sistemático. Esse protocolo estabelece critérios a serem seguidos para atividades específicas relacionadas à coleta, tratamento e armazenamento dos dados, resultando na criação de um repositório de dados local.

A proposição de um protocolo para a coleta e tratamento de dados de patentes representa uma abordagem metódica e organizada, destinada a executar um conjunto finito de atividades com o objetivo de recuperar e integrar dados de patentes provenientes de diferentes fontes. Tal abordagem, ainda pouco explorada na literatura, busca fornecer uma estrutura clara de procedimentos, aliada à utilização de ferramentas e métodos adequados.

Ademais, a proposição do protocolo proporciona uma estrutura clara para orientar a execução das atividades necessárias, reduzindo, assim, entraves que possam surgir durante o desenvolvimento de pesquisas envolvendo dados de patentes. Além disso, a

implementação de um protocolo constitui um ambiente evolutivo, passível de atualização e aprimoramentos mediante a identificação de problemas que possam emergir devido ao constante avanço tecnológico e atualizações dos repositórios de patentes.

Fundamentado no modelo genérico proposto por [Abbas, Zhang e Khan \(2014\)](#), este estudo amplia e complementa a abordagem dos autores ao detalhar os processos a serem executados nas camadas de pré-processamento e processamento.

## 4 Metodologia

A palavra método tem como origem a palavra grega *methodos* e é composta de *metá*, que quer dizer através de, por meio de e de *hodós* que significa via, caminho (RODRIGUES, 2006 apud ALMEIDA, 2016). Assim, a metodologia tem como objetivo estabelecer os métodos e caminhos utilizados para se atingir os objetivos estabelecidos nesta tese.

Este capítulo apresenta um roteiro sistemático para a investigação conduzida neste estudo, detalhando os métodos adotados. Fundamentado em princípios teóricos e metodológicos relevantes à área de pesquisa, o capítulo tem como objetivo orientar o leitor na compreensão da abordagem empregada, bem como possibilitar a replicação do estudo. O capítulo está organizado em quatro seções: questão de pesquisa, hipótese, estrutura metodológica e processos operacionais de pesquisa.

### 4.1 Questão de pesquisa

Evidenciada a importância do uso das informações contidas em documentos de patentes, o presente trabalho pretende responder a seguinte questão de pesquisa:

De que maneira a implementação de um protocolo sistemático de coleta pode mitigar as limitações de completude de dados inerentes aos repositórios de patentes de acesso aberto?

### 4.2 Hipótese

A criação de um protocolo de coleta, fundamentada em princípios de Ciência de Dados e Ciência da Informação, reduz significativamente a assimetria entre disponibilidade legal e acessibilidade funcional a dados bibliográficos de patentes, permitindo análises patentométricas de larga escala.

### 4.3 Estrutura referencial metodológica

Nesta seção, apresenta-se a estrutura metodológica de referência desta tese, destacando-se a abordagem, os métodos e as técnicas empregados na condução da pesquisa científica, à luz dos referenciais teóricos de Gil (1991), Lakatos e Marconi (1993), Rudio (2000), Silva e Menezes (2005).

A tese possui natureza de pesquisa aplicada, desenvolvida com orientação epistêmica hipotético dedutiva alinhada com o problema, objetivo, hipótese e questão central

anteriormente apresentados. O foco está em gerar conhecimento prático voltado para a resolução de problemas específicos no campo da propriedade industrial, particularmente na coleta e análise de dados de patentes em repositórios de acesso aberto.

Adicionalmente, a pesquisa adota uma orientação metodológica exploratória e descritiva, pois busca explorar novas estratégias e ferramentas para a coleta de dados, ao mesmo tempo em que descreve as barreiras e oportunidades relacionadas ao uso de dados de patentes. Também propõe metodologias para otimizar esse processo.

A abordagem adotada é tanto qualitativa quanto quantitativa, ou seja, uma abordagem quali-quantitativa, pois está centrada na produção e processamento de informações baseadas em dados numéricos e textuais.

O método utilizado é o hipotético-dedutivo, caracterizado pela formulação de hipóteses a partir de uma teoria existente, seguidas pela dedução de consequências testáveis, as quais são verificadas por meio da experimentação ou análise de 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 propriedade industrial, patentometria e ciência da informação.

As técnicas de pesquisa utilizadas nesta tese incluem levantamento documental e bibliográfico de informações, que podem estar disponíveis tanto em formato digital quanto impresso. As fontes de pesquisa abrangem dados provenientes de fontes primárias, secundárias e terciárias. As fontes primárias referem-se a dados gerados pelo autor ou por terceiros, como documentos oficiais, legislações e relatórios. As fontes secundárias resultam da compilação de dados primários, como livros, periódicos e repositórios de patentes. Já as fontes terciárias são constituídas pela combinação de dados primários e secundários, como catálogos de bibliotecas.

Em síntese, a estrutura referencial metodológica da tese está representada no [Quadro 4](#).

Quadro 4 – Referencial metodológico adotado na tese

| Aspectos metodológicos | Delimitação                             |
|------------------------|-----------------------------------------|
| Natureza               | Pesquisa Aplicada                       |
| Objetivo               | Exploratório e Descritivo               |
| Abordagem              | Qualitativa e Quantitativa              |
| Método                 | Hipotético-dedutivo                     |
| Procedimento           | Documental e Bibliográfica              |
| Técnicas               | Levantamento Documental e Bibliográfico |
| Fontes de informação   | Primária, secundária e terciária        |

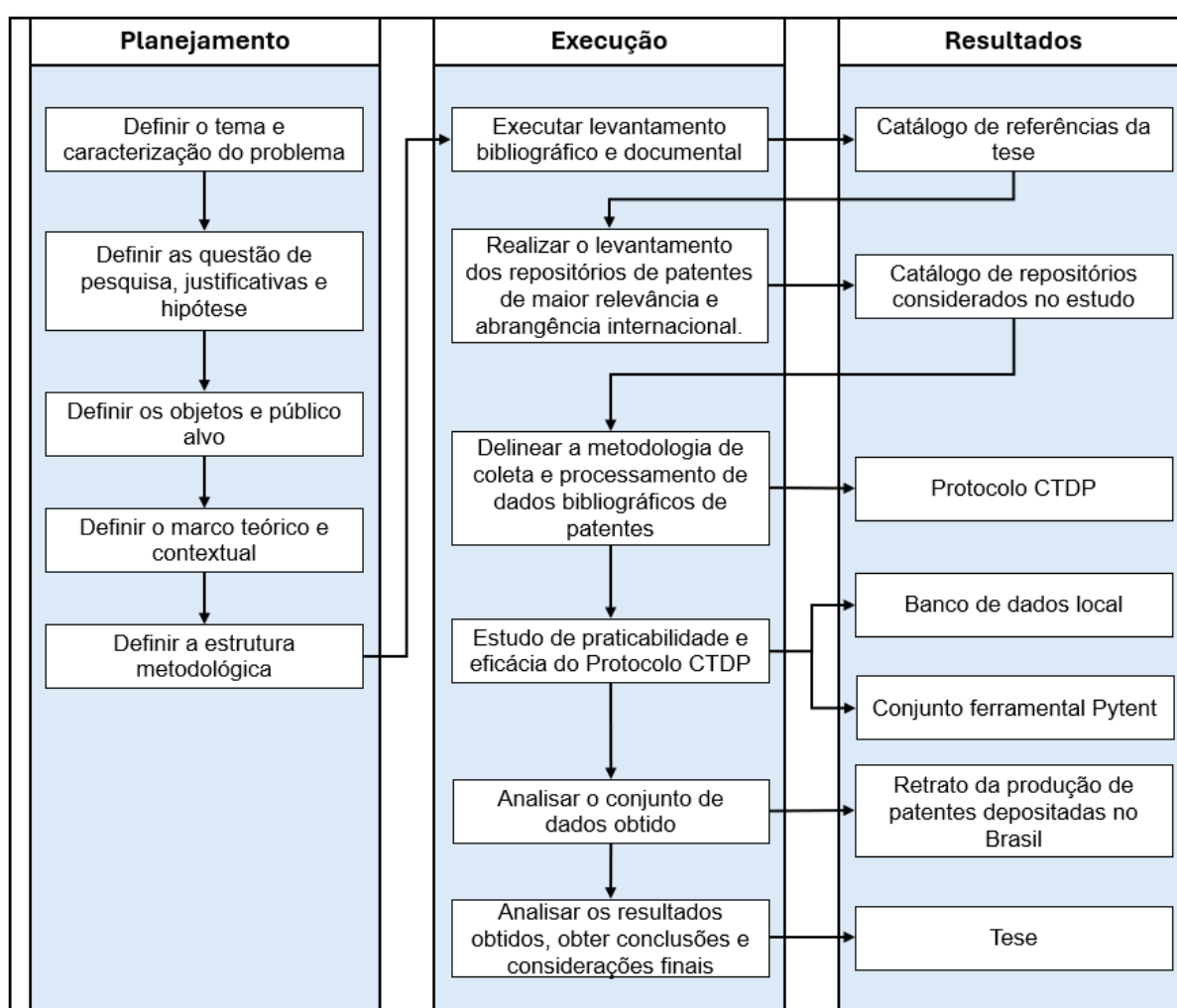
O [Quadro 4](#) evidencia que a tese empregará uma combinação de procedimentos e técnicas, com o objetivo de alcançar os resultados propostos a partir de diferentes tipos de

fontes de informação, todas estruturadas sob a metodologia hipotético-dedutiva.

## 4.4 Processos operacionais de pesquisa

Os processos operacionais desta tese foram estruturados em três estágios distintos, que se complementam de forma progressiva na construção do protocolo de coleta e tratamento de dados bibliográficos de patentes em repositórios de acesso aberto. Cada etapa se relaciona de acordo com o fluxo apresentado na [Figura 11](#).

Figura 11 – Esquema do fluxo dos processos operacionais de pesquisa.



Fonte: Elaboração do autor

O esquema apresentado na figura é composto por três conjuntos de atividades sequenciais, com o objetivo de delimitar o escopo da tese e garantir sua conclusão dentro do prazo estipulado pelo Programa de Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC) do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG).

No primeiro conjunto, composto por cinco blocos, concentram-se as atividades relacionadas à etapa de planejamento da tese. No primeiro bloco, ocorreu a definição do tema e dos elementos que caracterizam o problema abordado na pesquisa. Em seguida, com base no tema e na caracterização do problema, são delineados a questão de pesquisa, a justificativa e a hipótese. Posteriormente, estabelecem-se os objetivos e o público-alvo, seguidos pela definição do marco teórico e contextual da pesquisa.

O último bloco de atividades do planejamento abrange a definição da metodologia da tese, incluindo a escolha da abordagem, dos métodos, das técnicas de pesquisa e das fontes de informação.

Com a conclusão da etapa de planejamento, tem início a fase de execução da tese, composta por seis blocos de atividades que implementam o planejamento estabelecido. Essas atividades incluem o levantamento bibliográfico e documental, a identificação dos principais repositórios de patentes de abrangência internacional, a definição da metodologia de coleta e processamento de dados bibliográficos de patentes, a elaboração e execução do estudo de viabilidade da metodologia proposta, a análise do conjunto de dados obtido para sua caracterização e, por fim, a análise dos resultados para formular as conclusões e considerações finais da tese.

Finalmente, os resultados obtidos na etapa de execução compõem os blocos da última fase, que culmina na elaboração do texto da tese. Esta etapa envolve a análise dos resultados e a formulação das conclusões, onde se confrontam os objetivos propostos com os resultados alcançados, incluindo os desdobramentos e eventos ocorridos durante o desenvolvimento da pesquisa.

#### 4.4.1 Levantamento bibliográfico e documental

A técnica de levantamento bibliográfico e documental adotada nesta tese tem como finalidade a construção dos marcos teórico e contextual, que servem de base para a formulação/sustentação da questão de pesquisa, das hipóteses, dos objetivos e das justificativas do estudo. Além disso, essa técnica fundamenta a elaboração da metodologia de coleta e tratamento dos documentos de patentes.

Esta etapa envolve uma revisão abrangente e sistemática de fontes teóricas e empíricas relacionadas à coleta, tratamento e análise de documentos de patentes, considerando as áreas de Propriedade Industrial e Ciência da Informação. O processo inclui a identificação, coleta e análise de materiais acadêmicos, como livros, artigos, teses, relatórios e documentos oficiais, com o objetivo de contextualizar e fundamentar teoricamente a pesquisa, além de identificar lacunas no conhecimento existente. Adicionalmente, esta fase abrange o levantamento dos principais repositórios de patentes de cobertura internacional, bem como a avaliação de suas funcionalidades e limitações, visando subsidiar o desenvolvimento de

metodologias e ferramentas para facilitar o acesso e a análise desses documentos.

O **Quadro 5** apresenta um compilado das fontes adotadas no desenvolvimento deste estudo.

Quadro 5 – Compilado das fontes de pesquisas adotadas na tese.

| <b>Tipo</b>                                 | <b>Nome</b>                                                  | <b>Motivação</b>                                                                                                                                                                                                                                                                                                                                                    |
|---------------------------------------------|--------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Plataforma digital de literatura científica | Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) | Plataformas digitais que oferecem acesso a uma extensa coleção de literatura científica e técnica, incluindo artigos de periódicos, livros e outras publicações especializadas, abrangendo diversas áreas do conhecimento.                                                                                                                                          |
|                                             | Portal de Periódicos CAPES/MEC                               |                                                                                                                                                                                                                                                                                                                                                                     |
|                                             | SciELO                                                       |                                                                                                                                                                                                                                                                                                                                                                     |
|                                             | ScienceDirect                                                |                                                                                                                                                                                                                                                                                                                                                                     |
| Plataformas de busca                        | Google Acadêmico                                             | Plataformas de busca especializada que permitem acessar literatura acadêmica e científica, como artigos de periódicos, livros, teses, dissertações, conferências e relatórios técnicos.                                                                                                                                                                             |
|                                             | Microsoft Academic                                           |                                                                                                                                                                                                                                                                                                                                                                     |
| Portais institucionais                      | INPI                                                         | É uma autarquia federal, responsável pela administração do sistema de propriedade industrial no Brasil. Atua na proteção de direitos relacionados a patentes, entre outros. Sua principal função é garantir a propriedade intelectual, incentivando a inovação e o desenvolvimento tecnológico, além de promover a segurança jurídica para os direitos industriais. |
|                                             | OMPI                                                         | A OMPI trabalha para harmonizar e coordenar os sistemas nacionais de propriedade intelectual, garantindo que os direitos dos criadores sejam protegidos de forma eficiente e eficaz em todo o mundo.                                                                                                                                                                |

Conforme apresentado, foram adotadas oito fontes de busca para a tese, organizadas em três categorias: plataformas digitais de literatura científica, plataformas de busca e portais institucionais. A prospecção bibliográfica foi conduzida por meio de uma estratégia de busca estruturada para garantir a replicabilidade do levantamento.

A construção das expressões de busca fundamentou-se no cruzamento de descritores em português e inglês, utilizando operadores booleanos (AND e OR) para refinar a recuperação. Os termos foram organizados em três eixos semânticos: o objeto de estudo (patente/patent), a métrica aplicada (patentometria/patentometrics) e o domínio técnico (coleta de dados/data collection; repositórios/repositories).

Desta forma, as seguintes expressões foram adotadas:

(patent\* OR “industrial property” OR “propriedade industrial”) AND (patentometr\* OR bibliometr\*) AND (“data collection” OR “coleta de dados” OR “metadata” OR “metadados” OR “open access” OR “acesso aberto”)

uma expressão mais abrangente, com foco em metodologia e coleta. E uma expressão mais específica

(patent\* OR “propriedade industrial”) AND (“open access” OR “acesso aberto”) AND (repository OR repositório) AND (“data mining” OR “coleta” OR “processamento”)

com foco em repositórios e desafios de dados.

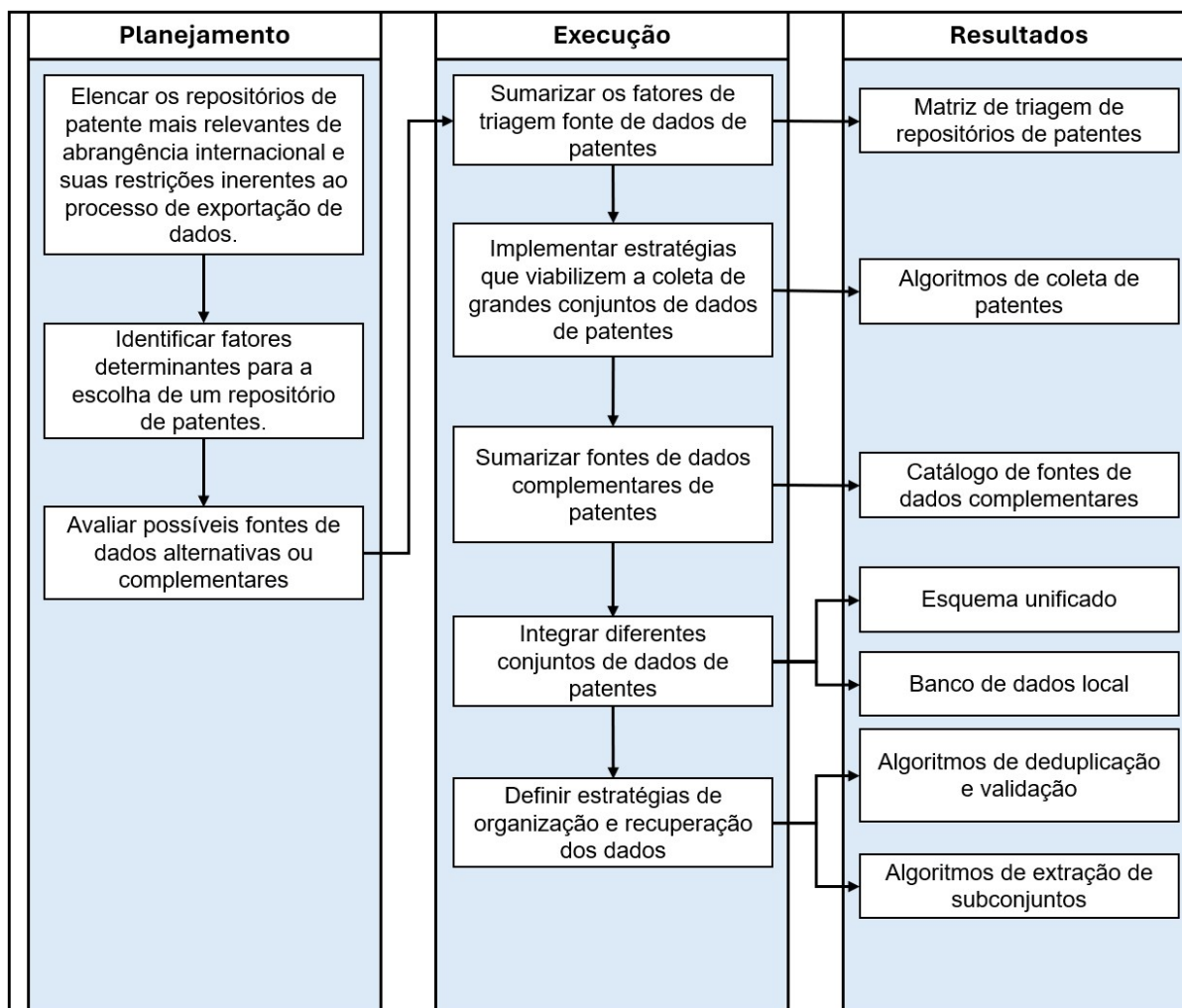
Para refinar a busca, sempre que a quantidade de resultados superasse cinquenta, foi aplicada uma restrição temporal, limitando a pesquisa a publicações dos últimos cinco anos. Com base nos resultados obtidos, seguindo a estratégia de busca estabelecida, foi realizada a seleção de documentos considerados relevantes para a tese. O método de seleção consistiu na leitura dos títulos e resumos para verificar a pertinência em relação à questão de pesquisa e aos objetivos da tese. Os documentos pertinentes constituíram o catálogo de referência da tese, os demais documentos foram desconsiderados.

#### 4.4.2 Construção da metodologia de coleta e tratamento dos dados de patentes

A metodologia de coleta e tratamento de dados bibliográficos de patentes, denominada Protocolo de Coleta e Tratamento de Dados de Patentes (Protocolo CTDp), foi desenvolvida com base nos insumos obtidos a partir do levantamento bibliográfico e documental. Esse levantamento permitiu identificar os principais desafios enfrentados pelos pesquisadores ao utilizar patentes como fonte de pesquisa, orientando o desenvolvimento de uma abordagem sistemática para superar essas dificuldades.

A metodologia proposta nesta tese é baseada no modelo genérico desenvolvido por [Abbas, Zhang e Khan \(2014\)](#). O processo de concepção da metodologia proposta é ilustrado na [Figura 12](#), cujo detalhamento será apresentado a seguir.

Figura 12 – Fluxo operacional para a construção da metodologia de coleta e tratamento de dados de patentes.



Fonte: Elaboração do autor

A figura apresentada representa o fluxo operacional metodológico adotado para a construção do protocolo de coleta e tratamento de dados bibliográficos de patentes, dividida em três fases principais: Planejamento, Execução e Resultados. Cada fase possui etapas específicas que descrevem atividades e processos necessários para o desenvolvimento do protocolo, com o objetivo de sistematizar a coleta e o tratamento de dados de patentes.

Na fase de planejamento, três atividades principais são realizadas. O primeiro bloco de atividade consiste em identificar repositórios de patentes com cobertura internacional, assim como a compreensão de restrições associadas ao processo de exportação dos dados adotados por estes repositórios. Em síntese, o objetivo é identificar as principais bases de dados de patentes com cobertura internacional, verificando suas limitações e restrições de acesso ou exportação de dados, como limitações de download ou acesso pago.

Em sequência, o próximo bloco de atividade é caracterizado por descrever os fatores

determinantes para a escolha de um repositório de patentes. Esta etapa foca em analisar critérios específicos que guiam a seleção de um repositório de patentes, como a relevância da base de dados, o volume de informações disponíveis e a qualidade dos metadados.

E, por fim, a última atividade compreende avaliar possíveis fontes de dados alternativas ou complementares. Envolve a busca por fontes complementares que podem fornecer dados adicionais às bases de patentes tradicionais, enriquecendo as análises com dados mais detalhados ou específicos.

Concluídas as atividades de planejamento, inicia-se a fase de execução. A execução envolve uma série de cinco atividades operacionais. A primeira atividade abrange sumarizar os fatores de triagem de fontes de dados de patentes identificados na fase de planejamento. Em outras palavras, esta etapa consiste em analisar e agrupar os critérios semelhantes com o objetivo de sintetizar e estabelecer um balizador que sirva de parâmetro para apoiar a escolha de um determinado repositório de patentes como fonte de coleta.

Em seguida, a segunda atividade consiste na implementação de estratégias que possibilitem a coleta de grandes volumes de dados de patentes no repositório selecionado na etapa anterior. Nesta fase, considera-se um "grande volume de dados" quando a coleta não pode ser concluída em um único processo. Assim, com base em uma análise detalhada dos critérios de exportação de dados definidos pelo repositório, são desenvolvidas e aplicadas técnicas e ferramentas que permitam coletas incrementais até que todos os dados estabelecidos sejam obtidos, sempre em conformidade com as diretrizes do repositório de patentes.

Dando continuidade, a terceira atividade da fase de execução consiste em resumir as fontes de dados complementares de patentes identificadas na etapa de planejamento. Esta atividade envolve a definição dos processos de coleta e organização dessas fontes de dados complementares mais relevantes.

A quarta atividade consiste em mapear a estrutura dos diversos conjuntos de dados coletados, com o objetivo de definir o esquema de organização de cada um deles. Depois disso, analisar e estabelecer diretrizes para construir um esquema único que viabilize implementar um banco de dados integrado, capaz de unificar e consolidar os diferentes conjuntos de dados em uma base adequada para análise subsequente.

A última atividade da fase de execução consiste em definir estratégias de organização e recuperação dos dados. Nesse estágio, são estabelecidas as diretrizes para estruturar os dados de acordo com o esquema elaborado na atividade anterior, bem como os métodos para sua recuperação dos dados, de modo a permitir a extração eficiente das informações. Esse processo inclui a resolução de problemas relacionados a redundâncias, inconsistências e integridade.

Por fim, a fase de resultados apresenta os produtos gerados a partir das etapas de

planejamento e execução. A “Matriz de Triagem de Repositórios de Patente” consiste em um documento que lista e categoriza os repositórios de patentes mais relevantes, detalhando suas restrições e vantagens em conformidade com o escopo de pesquisa. Os “Algoritmos de Coleta de Patentes” são ferramentas desenvolvidas para automatizar a coleta de dados em diferentes repositórios, facilitando o acesso a conjunto de patentes. O “Catálogo de Fontes de Dados Complementares” organiza e descreve essas fontes, indicando suas características e formas de integração ao conjunto de dados principal.

O “Esquema Unificado” representa a estrutura de integração dos diversos conjuntos de dados sobre patentes, fornecendo uma base para a criação de um banco de dados consolidado. Esse banco de dados local, denominado “Banco de Dados Local”, é construído a partir da integração dos dados conforme o esquema unificado, permitindo a gestão centralizada das informações.

Para assegurar a qualidade dos dados, foram desenvolvidos “Algoritmos de Deduplicação e Validação”, que têm como função remover duplicatas e validar as informações coletadas. Finalmente, os “Algoritmos de Extração de Subconjuntos” permitem a segmentação dos dados, facilitando a extração de subconjuntos específicos de patentes e possibilitando análises mais detalhadas e direcionadas.

A partir do delineamento metodológico apresentado, o próximo capítulo detalha o desenvolvimento da tese

## 5 Desenvolvimento

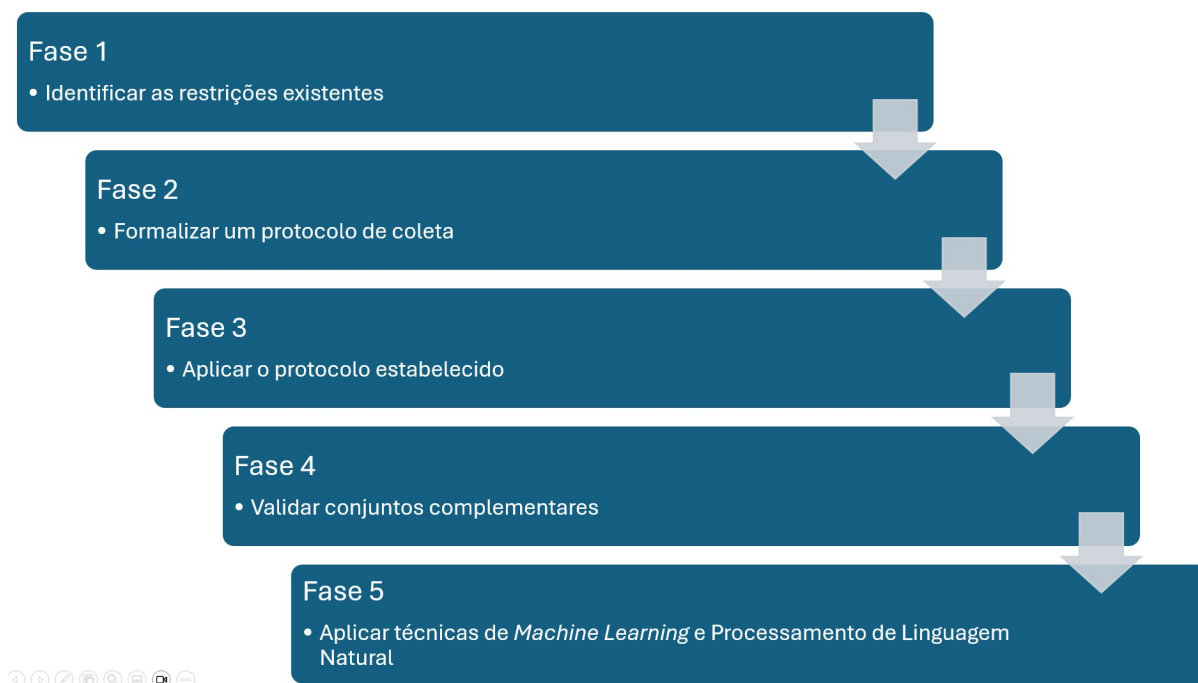
Este capítulo descreve detalhadamente o processo de desenvolvimento da metodologia proposta para a coleta e tratamento de dados bibliográficos de patentes. A partir dos marcos teóricos e do levantamento bibliográfico realizado, foram identificados os principais desafios e lacunas que fundamentam a necessidade de uma abordagem estruturada e sistemática para o tratamento de grandes volumes de dados de patentes.

Inicialmente, são apresentados os critérios e diretrizes utilizados para a seleção das fontes de dados e repositórios de patentes, bem como as especificidades de cada um em relação à abrangência geográfica e temporal, restrições de acesso e formatos de disponibilização dos dados. Em seguida, detalham-se as estratégias e ferramentas desenvolvidas para a coleta incremental dos dados, considerando as limitações impostas por cada repositório e os requisitos de consistência e integridade da base de dados consolidada.

Para a integração dos diferentes conjuntos de dados coletados, este capítulo aborda a construção de um esquema unificado, que permite a harmonização das informações e a eliminação de redundâncias. Além disso, são descritos os algoritmos desenvolvidos para a deduplicação, validação e extração de subconjuntos específicos de patentes, visando a flexibilidade e a eficiência na recuperação dos dados para análises posteriores.

Por meio dessa abordagem, o capítulo visa descrever o processo de desenvolvimento da metodologia, além de demonstrar como cada etapa foi concebida para responder aos objetivos específicos da pesquisa. A [Figura 13](#) apresenta uma visualização gráfica de como está organizado o capítulo de desenvolvimento.

Figura 13 – Visão geral do capítulo de desenvolvimento



Fonte: Elaboração do autor

O capítulo de desenvolvimento deste estudo é composto por cinco fases incrementais, como detalhado na [Figura 13](#). A Fase 1 caracteriza as restrições de acesso e limitações técnicas dos repositórios de acesso abertos ([Seção 5.1](#)). Na Fase 2, procede-se à formalização do protocolo sistemático de coleta, definindo-se a lógica de extração, normalização e armazenamento dos dados ([Seção 5.2](#)). A Fase 3 compreende a aplicação prática do protocolo em larga escala ([Seção 5.3](#)), enquanto a Fase 4 foca na validação de conjuntos de dados complementares oriundos da Plataforma Lattes ([Seção 5.4](#)). Finalmente, a Fase 5 apresenta o emprego de técnicas avançadas de *Machine Learning* e Processamento de Linguagem Natural para a categorização semântica dos dados ([Seção 5.5](#)).

## 5.1 Avaliação das restrições existentes no processo de coleta de patentes

Esta investigação teve como ênfase identificar os repositórios mais relevantes que oferecem acesso aberto a dados bibliográficos de patentes, além de compreender as restrições inerentes à recuperação dessas informações. Ressalta-se que a pesquisa concentrou-se em repositórios de abrangência internacional, ou seja, aqueles que disponibilizam dados provenientes de múltiplos escritórios de patentes, de modo a assegurar maior cobertura e representatividade no contexto global de informação tecnológica.

A partir dos insumos coletados na etapa de levantamento bibliográfico e documental

da tese, e em concordância com as recomendações de [Brasil \(2015\)](#), foram identificados os repositórios mais relevantes para esta análise: Espacenet, Latipat, Patentscope, Google Patents, Lens, Derwent Innovation Index e Orbit Intelligence.

Para explorar outras fontes de dados além dos documentos científicos, utilizou-se a busca avançada da ferramenta Google<sup>1</sup>, com o intuito de identificar repositórios de patentes que ofereçam acesso aberto a dados bibliográficos e que sejam distintos dos amplamente utilizados na literatura, como Espacenet, Latipat e Patentscope. Assim, foi desenvolvida a sentença de busca `Patent* search -espacenet -latipat -patentscope`, que permite localizar resultados contendo a palavra “patent” em suas diversas variações (como **patente**, **patentometria** e **patentometry**), incluindo o termo **search** e excluindo as menções a **espacenet**, **latipat** e **patentscope**.

A pesquisa foi realizada em fevereiro de 2024, seguindo a estratégia metodológica previamente definida. Delimitou-se o período de análise entre os anos de 2013 e 2023, com o intuito de contemplar apenas os registros mais recentes. Essa delimitação temporal visa assegurar que os dados analisados refletem um panorama mais atual possível para este estudo, garantindo, assim, a relevância e a aplicabilidade dos resultados obtidos.

Dessa forma, apenas os 50 primeiros resultados foram considerados na análise, uma vez que são os mais relevantes de acordo com os critérios de busca. A classificação desses resultados segue o sistema de ranqueamento adotado pela ferramenta de busca Google, reconhecida atualmente como o mecanismo de busca líder em sua categoria. O Google emprega um sistema complexo de classificação que envolve múltiplas etapas de coleta, indexação e ranqueamento dos dados. Para uma visão detalhada sobre o sistema de classificação de resultados do Google, é possível acessar o endereço eletrônico: <https://developers.google.com/search/docs/appearance/ranking-systems-guide?hl=pt-br>. Os resultados considerados para a pesquisa estão listados no [Apêndice A](#).

Os 50 primeiros resultados, portanto, foram classificados como os mais relevantes para o processo de análise. Cada resultado corresponde a uma página web mencionando dados de patentes. A análise, de abordagem dedutiva, consistiu em acessar individualmente cada página pelos respectivos links, investigando a presença de ferramentas para pesquisa de patentes. Com base nesse acesso, os resultados foram organizados em três categorias: (1) páginas que fazem referência a conjuntos específicos de patentes, (2) páginas com ferramentas de análise e busca de patentes, e (3) páginas de repositórios de escritórios regionais de patentes, oferecendo acesso a dados de apenas um escritório específico. Para a próxima etapa da análise, apenas as páginas da categoria (2) foram consideradas, conforme descrito no [Quadro 6](#).

Entre as ferramentas listadas no quadro, que oferecem recursos para pesquisa

<sup>1</sup> [https://www.google.com.br/advanced\\_search](https://www.google.com.br/advanced_search)

Quadro 6 – Grupo (02) ferramentas de análise e/ou de busca de patentes

| <b>Ferramentas</b>                                          | <b>Descrição</b>                                                                                                                                                                                                                             |
|-------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| The Lens                                                    | É uma ferramenta fácil que oferece vários recursos de pesquisa, incluindo pesquisa normal, pesquisa estruturada, busca biológica e busca de classificação de patentes. Em um banco de dados de mais de 125 milhões de registros de patentes. |
| Google Patents                                              | O Google Patentes inclui mais de 120 milhões de publicações de patentes de mais de 100 escritórios de patentes em todo o mundo, bem como muitos outros documentos técnicos e livros indexados no Google Acadêmico e no Google Livros.        |
| Patent Search Service - Questel                             | O Orbit Intelligence, um software dedicado à análise de patentes e à pesquisa.                                                                                                                                                               |
| CAS PatentPak® - Full-Text Patent Search                    | É uma solução de fluxo de trabalho integrada para a pesquisa de patentes sobre química. Oferece acesso a mais de 18 milhões de patentes de 46 grandes escritórios de patentes em todo o mundo.                                               |
| Patentfield   AI Patent Search, Analytics and investigating | Uma plataforma de busca e análise de patentes que combinada a Pesquisa assistida, visualização de dados, IA, entre outros.                                                                                                                   |
| The Global Patent Search Tool for IP Users of All Levels    | Ferramenta de busca de patentes                                                                                                                                                                                                              |
| Derwent Innovation - Patent Search for IP Professionals     | Derwent Innovations Index, pesquisa rápida e precisa de patentes e citações de invenções em produtos químicos.                                                                                                                               |

de documentos de patentes, foi investigada a disponibilidade de acesso gratuito a estas ferramentas, bem como a extensão da cobertura de patentes, ou seja, a quantidade de escritórios de patentes disponíveis para pesquisa. Com base nesses critérios, apenas os repositórios The Lens e Google Patents foram selecionados; no entanto, ambos já compõem o conjunto de repositórios de patentes considerados neste estudo.

Portanto, com base nesta análise preliminar, os repositórios de patentes adotados neste estudo são: Espacenet, Google Patents, Latipat, Patentscope e The Lens. Após a definição do conjunto de repositórios, realizou-se uma análise das restrições associadas à recuperação de dados em cada um deles. Esse processo foi conduzido por meio do acesso às páginas de ajuda e sessões informativas de cada repositório, visando compreender as especificidades do processo de coleta de dados. Observou-se que cada repositório possui particularidades próprias. A [Tabela 2](#) apresenta um resumo das principais características relativas à recuperação de dados em cada repositório avaliado.

Tabela 2 – Restrições inerentes à recuperação de dados de patentes nos repositórios analisados neste estudo

| Repositório    | Acesso | Tamanho da base de dados            | Limite exportação | Formato exportação     |
|----------------|--------|-------------------------------------|-------------------|------------------------|
| Espacenet      | Aberto | 149 milhões de documentos           | 500 registros     | XLS, CSV               |
| Google Patents | Aberto | 120 milhões de documentos           | 1.000 registros   | CSV, XLSX              |
| Latipat        | Aberto | Utilizada a base de dados Espacenet | 500 registros     | XLS, CSV               |
| Lens           | Aberto | 133 milhões de documentos           | 1.000 registros   | CSV, RIS, BibTeX, JSON |
| Patentscope    | Aberto | 115 milhões de documentos           | —                 | —                      |

A tabela apresentada resume as principais características e limitações dos repositórios de patentes considerados na pesquisa, especialmente no que se refere ao acesso, ao tamanho da base de dados, ao limite de registros para exportação, e aos formatos disponíveis para exportação. A coluna “Acesso” indica que todos os repositórios analisados oferecem acesso aberto ao público, permitindo que usuários consultem dados de patentes sem restrições de assinatura ou pagamento. A coluna “Tamanho da base de dados”, apresenta o número estimado de documentos de patentes disponíveis em cada repositório. Destaca-se o Espacenet com aproximadamente 149 milhões de documentos, seguido pelo Lens (133 milhões) e Google Patents (120 milhões), enquanto Patentscope disponibiliza cerca de 115 milhões de registros. O repositório Latipat é baseado na base de dados do Espacenet, o que indica que seu tamanho de base reflete os dados disponíveis neste último.

Já em relação aos limites de exportação, a coluna “Limite de exportação” revela o número máximo de registros que cada repositório permite exportar em uma única operação. Espacenet e Latipat limitam a exportação a 500 registros, enquanto Google Patents e Lens permitem a exportação de até 1.000 registros. O Patentscope não explicita limites formais para a exportação de dados; entretanto, o acesso a conjuntos mais abrangentes de informações é disponibilizado mediante pagamento. Contudo, é importante destacar que o repositório Lens, mediante a criação de uma conta gratuita, permite a exportação de 50.000 registros.

A última coluna, “Formato de exportação”, relaciona os formatos de exportação de dados disponibilizados por cada repositório. Espacenet e Latipat oferecem exportação em formatos XLS e CSV, Google Patents permite exportação em CSV e XLSX, e Lens apresenta uma gama mais ampla de formatos, incluindo CSV, RIS, BibTeX e JSON. Patentscope, por outro lado, não fornece informação sobre formatos de exportação.

Este comparativo fornece uma visão objetiva sobre as restrições e potencialidades de cada repositório sob a perspectiva de coleta de dados de patentes. Assim, a análise evidencia quatro obstáculos principais que limitam uma coleta mais abrangente desses

dados, os quais podem ser resumidos nos seguintes fatores:

- **FATOR01:** Limite de Exportação de Dados. A restrição no número de registros que podem ser exportados de uma vez em repositórios como Espacenet e Latipat (500 registros) e, em menor grau, Google Patents e Lens (1.000 registros), representa uma barreira significativa.
- **FATOR02:** Coletas Incrementais. O limite de exportação de dados exige que o pesquisador realize múltiplas operações para coletar grandes volumes de dados, o que demanda mais tempo e aumenta a complexidade do processo de coleta.
- **FATOR03:** Formatos de Exportação. Os formatos de exportação limitam a flexibilidade de manipulação e relacionamento dos dados coletados, principalmente em análises que necessitam de cruzamento de dados.
- **FATOR04:** Cobertura e Tamanho da Base de Dados. A variação no tamanho e cobertura das bases de dados limita a abrangência da coleta. Alguns repositórios, como Patentscope e Latipat, não oferecem cobertura tão extensa quanto o Espacenet ou o Lens, o que pode resultar em dados menos representativos para estudos que necessitam de uma visão mais completa e global.

Com base nos fatores identificados, este estudo delineou estratégias que se destacam como possíveis soluções para os desafios impostos por cada um dos fatores mencionados. Os detalhes dessas soluções encontram-se descritos no [Quadro 7](#).

Quadro 7 – Estratégias que objetivam a solução dos desafios encontrados.

| Desafio | Solução proposta                                                                            |
|---------|---------------------------------------------------------------------------------------------|
| FATOR01 | (1) Estratégia de paginação; (2) Alternar entre repositórios; (3) Uso de APIs avançadas.    |
| FATOR02 | (1) Coleta incremental baseada em data de publicação; (2) Automatizar o processo de coleta. |
| FATOR03 | (1) Converter formatos e padronizar os dados; (2) Criar um Pipeline de integração.          |
| FATOR04 | (1) Processo de triagem de repositórios de patentes; (2) Fontes Complementares de dados.    |

As estratégias apresentadas pelo quadro supracitado, combinadas com ferramentas de automação e processamento de dados, podem auxiliar na coleta de dados de patentes de maneira mais eficiente, respeitando todas as diretrizes de coleta impostas pelos repositórios.

Para solucionar as limitações descritas no FATOR01, são adotadas três procedimentos: (1) Explorar as opções de filtros disponibilizados pelos repositórios, com objetivo de realizar consultas incrementais que viabilizem a recuperação de todo o conjunto de interesse. Inclusive, avaliar a existência de recursos avançados como a aplicação de instruções CQL (*Contextual Query Language*) para contribuir na elaboração de expressão de buscas mais precisas; (2) Utilizar múltiplos repositórios e explorar fontes complementares com menor limite de exportação, integrando os dados coletados; (3) Quando disponíveis pelo repositório, fazer uso de APIs. As APIs geralmente têm permissões de exportação

mais amplas, permitindo que volumes maiores de dados sejam acessados em uma única operação ou com limites menos restritivos.

Já em relação às limitações referentes ao FATOR02, são empregados dois procedimentos: (1) Filtrar dados de acordo com o intervalo de datas de publicação dos registros permite realizar coletas periódicas de dados novos sem duplicar os registros já armazenados. (2) Desenvolver scripts que realizam coletas automáticas em intervalos específicos, controlando logs e identificadores únicos para evitar duplicidades e garantir que todos os registros sejam coletados ao longo do tempo.

A abordagem empregada para o FATOR03 é composta por dois procedimentos: (1) Após a coleta, converter os dados exportados em formatos mais flexíveis que viabilizem a integração e relacionamento dos dados, usando scripts de conversão e padronização. (2) Criar um pipeline de ETL (Extração, Transformação e Carga) para processar e integrar dados de diferentes formatos de maneira padronizada.

E, por fim, em relação ao FATOR04, são aplicadas dois procedimentos: (1) Formalizar uma triagem de seleção de banco de dados a ser utilizado no processo de coleta. (2) Fazer uso de fontes de dados complementares, como bancos de dados abertos e regionais, considerando também publicações científicas ou bases de dados industriais que contenham dados de patentes mais específicos para uma cobertura mais ampla.

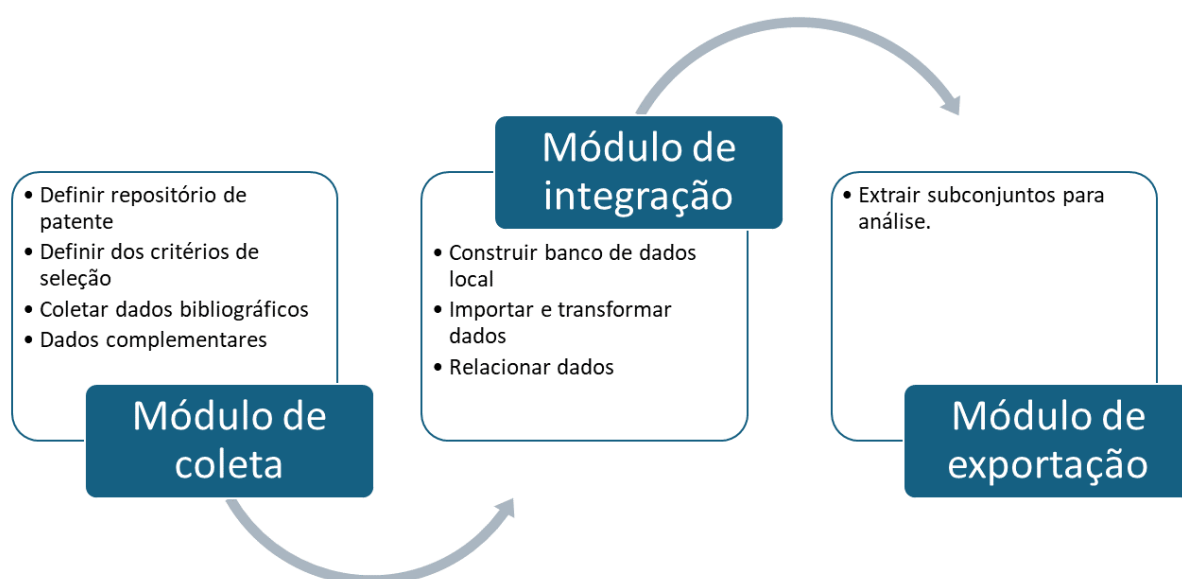
A seção a seguir detalha a implementação das estratégias propostas neste estudo e examina a relação entre os diferentes procedimentos apontados.

## 5.2 Protocolo de coleta e tratamento de dados bibliográficos de patentes

Para viabilizar a implementação das estratégias expostas anteriormente, foram definidos os requisitos necessários para a elaboração de um protocolo específico para a coleta de dados bibliográficos de patentes em repositórios de acesso aberto. O objetivo desse protocolo é estabelecer um processo sistemático que permita a coleta desses dados, oferecendo uma alternativa viável para superar os obstáculos identificados ao longo do estudo.

De encontro ao Objetivo A (Desenvolver um protocolo para a coleta de dados bibliográficos de patentes em repositórios de acesso aberto, de forma a viabilizar a extração de grandes conjuntos de patentes, em conformidade com diretrizes o repositório) o protocolo proposto é estruturado em módulos, concebidos para execução sequencial, mas que também podem ser executados de forma independente, conforme a necessidade. A comunicação entre os módulos sugeridos é representada de forma esquemática na [Figura 14](#).

Figura 14 – Esquema de relacionamento entre os módulos



Fonte: Elaboração do autor

O módulo de coleta é responsável por executar o processo de coleta de dados bibliográficos de patentes. Para isso, alguns requisitos precisam ser atendidos. O primeiro requisito é a definição do repositório de patentes a ser utilizado, uma vez que os campos aplicados como critérios de seleção e os recursos de coleta podem variar conforme o repositório escolhido. Logo, a escolha do repositório afeta diretamente o andamento de toda a pesquisa. Assim, a escolha do repositório é fundamentada em um sistema de triagem que serve como balizador para selecionar a opção mais adequada. O processo de triagem de repositórios de patentes é detalhado na [Subseção 5.2.1](#).

Após a seleção do repositório, é necessário definir quais campos serão aplicados como critérios de seleção, tais como: depositante, inventor, classificação, entre outros, de acordo com os recursos disponíveis no repositório escolhido. Para além dos filtros elencados, é necessário incluir um filtro por período de publicação, o que viabiliza a coleta incremental dos dados.

Dessa forma, o processo de coleta das patentes pode ser iniciado por meio de um sistema computacional que automatiza a coleta incremental de todas as patentes que atendam aos critérios estabelecidos, respeitando as restrições do repositório. O detalhamento desse processo encontra-se na [Subseção 5.2.3](#). As patentes coletadas são armazenadas localmente em arquivos, de acordo com o formato de exportação disponibilizado pelo repositório utilizado.

Adicionalmente, podem ser coletados dados complementares relacionados aos

registros de depósito de patentes armazenados. Entre esses dados, incluem-se relatórios técnicos disponibilizados por escritórios de patentes e currículos da Plataforma Lattes, com informações sobre depósitos de patentes. Esses dados, ainda que abranjam um grupo restrito, fornecem informações relevantes sobre os proponentes que não estão disponíveis nos dados bibliográficos das patentes como: formação acadêmica, endereço profissional, área de pesquisa e atuação, nacionalidade entre outras.

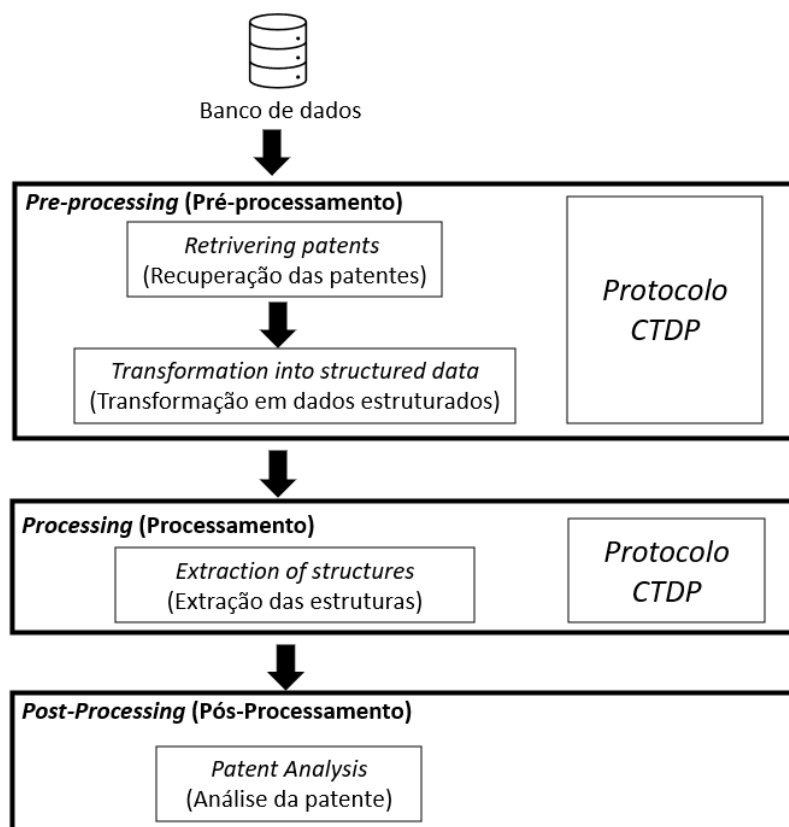
O módulo de integração tem como objetivo correlacionar os diferentes conjuntos de dados coletados, transformando-os em entidades bem definidas. Essas entidades são organizadas para estabelecer relações entre si, formando uma estrutura para consultas e cruzamento de dados. A base de dados relacional construída organiza as informações em tabelas, onde cada tabela representa uma entidade ou conceito específico. Cada linha de uma tabela corresponde a uma entrada de dados (um registro), enquanto as colunas representam os diferentes tipos de informações armazenadas para cada registro. Assim, o módulo de integração facilita a consolidação dos dados de várias fontes em um único conjunto integrado, proporcionando uma visão abrangente e detalhada das informações disponíveis.

A integração desses dados é essencial, pois permite reunir informações de diferentes fontes e formatos em uma única base, possibilitando o cruzamento de dados provenientes de diversas origens. Esse processo também facilita a identificação e correção de inconsistências e duplicidades.

Por fim, o módulo de exportação permite realizar consultas estruturadas na base de dados, possibilitando a extração de subconjuntos específicos de informações. Utilizando uma linguagem de consulta estruturada (SQL), este módulo permite a obtenção precisa dos dados necessários para gerar relatórios e fontes de análise, além de possibilitar a integração com ferramentas de análise de dados.

Os módulos e suas inter-relações propostas neste estudo constituem um protocolo sistemático para coleta, tratamento e armazenamento de dados de patentes, operando nas camadas de Pré-processamento e Processamento do modelo apresentado por [Abbas, Zhang e Khan \(2014\)](#) ([Figura 15](#)).

Figura 15 – Protocolo CTDp no processo genérico



Fonte: Elaboração do autor, adaptado de (ABBAS; ZHANG; KHAN, 2014)

Com o intuito de facilitar a identificação, comunicação e compreensão da solução aqui desenvolvida, este protocolo será referido como "Protocolo de Coleta e Tratamento de Dados Patentários" ou, de forma abreviada, "Protocolo CTDp".

Conforme ilustrado na [Figura 15](#), o Protocolo CTDp abrange a camada de Pré-processamento do modelo proposto por Abbas, Zhang e Khan, aplicando técnicas para recuperação de patentes e transformação dos dados em estruturas organizadas. Essas técnicas incluem a extração automática de informações e sua organização em formatos adequados para análises subsequentes. O protocolo também compreende a camada de Processamento, incorporando métodos para extração de conjuntos de dados e integração com ferramentas de análise e mineração de dados.

As próximas subseções apresentam uma descrição detalhada da estrutura completa do Protocolo CTDp.

### 5.2.1 Triagem de repositório de dados de patentes e fontes de dados complementares

No módulo de coleta, o processo de triagem de repositórios de patentes envolve a definição do repositório que será utilizado como fonte de dados. O acesso a esses conteúdos, por meio de bases de dados eletrônicas, facilita significativamente a pesquisa e o uso das informações contidas nas patentes. No entanto, cada repositório apresenta particularidades específicas, como o tamanho da base de dados, limites e formatos de exportação, além de dados adicionais, como procedimentos legais do processo de depósito e citações recebidas.

A escolha do repositório de patentes adequado é, portanto, um elemento essencial para o sucesso do processo de coleta de dados. De encontro ao Objetivo B (definir critérios objetivos para definir a seleção de repositórios para a coleta de dados bibliográficos de patentes), com base no levantamento bibliográfico e documental realizado neste estudo, foram estabelecidos 11 critérios, descritos no [Quadro 8](#), que servem como diretrizes para selecionar o repositório de patentes mais adequado como fonte de dados.

Quadro 8 – Critérios de seleção

| Nº | Critério                  | Descrição                                                                                                                                                                                                                                         |
|----|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 01 | Cobertura Geográfica      | Avaliar se o repositório abrange as patentes de todas as regiões geográficas de interesse da pesquisa. Como, por exemplo, patentes depositadas no Brasil.                                                                                         |
| 02 | Amplitudes de Documentos  | Avaliar se entre a variedade de tipos de documentos disponibilizado pelo o repositório, (como patentes de invenção, modelos de utilidade, design, famílias, citações, entre outros) é condizentes com o escopo da pesquisa.                       |
| 03 | Áreas Especializadas      | Avaliar se o repositório cobre os setores industriais específicos de interesse de pesquisa. Alguns repositórios de dados de patentes são especializadas em áreas específicas como química, biotecnologia, farmacêutica, entre outras.             |
| 04 | Atualização e Manutenção  | Avaliar a frequência que os dados do repositório são atualizados. Informações desatualizadas podem prejudicar a precisão das análises.                                                                                                            |
| 05 | Interface de Busca        | Avaliar a usabilidade da interface de busca. Uma interface intuitiva, com muitos recursos de filtros e seleção, pode facilitar o processo de pesquisa de documentos de patentes.                                                                  |
| 06 | Recursos de Análise       | Avaliar se o repositório disponibiliza recursos para análise sem a necessidade de realizar download dos dados. Ferramentas de análise avançadas podem otimizar o processo, eliminando o tempo de extração das informações para análise posterior. |
| 07 | Idioma                    | Avaliar se o repositório disponibiliza as patentes nos idiomas relevantes para pesquisa. O idioma pode interferir diretamente em análises textuais, como identificação de termos frequentes e remoção de <i>stopwords</i> .                       |
| 08 | Custo                     | Avaliar o investimento financeiro necessário para acesso ao repositório de dados de patentes. Alguns repositórios são de acesso gratuito, enquanto outros exigem assinatura ou pagamento por acesso.                                              |
| 09 | Relevância do Repositório | Avaliar a relevância e reputação do repositório de dados, bem como avaliações de outros pesquisadores. Considerar as pesquisas que usaram o repositório, bem como as limitações encontradas.                                                      |
| 10 | Exportação de Dados       | Avaliar se o repositório disponibiliza recursos para viabilizar exportação dos dados, como <i>download</i> nos formatos CSV, XLS, bem como a existência de possíveis limitações quanto a quantidade de registros que podem ser exportados.        |
| 11 | API                       | Avaliar se o repositório disponibiliza de recursos tecnológicos, como API ou <i>Webservices</i> para viabilizar a construção de processos computacionais de coleta de dados.                                                                      |

Os critérios apresentados são fundamentais para tornar o processo de escolha do repositório de patentes mais objetivo, ao estabelecer parâmetros claros e precisos para a seleção. Esses critérios visam assegurar a qualidade do conjunto de dados a ser utilizado na pesquisa, garantindo que o repositório selecionado atenda aos requisitos essenciais definidos pelo estudo.

Nesse contexto, a avaliação da cobertura geográfica (Critério 01) é essencial para garantir que o repositório inclua todas as regiões relevantes para a pesquisa, especialmente em estudos focados em patentes de países específicos ou blocos econômicos. A escolha de um repositório com cobertura geográfica adequada assegura que a pesquisa seja representativa e contextualizada nos mercados e jurisdições de interesse.

De igual modo, avaliar a amplitude de documentos (Critério 02) assegura que o repositório inclua uma variedade de tipos de documentos de patentes, como invenções e modelos de utilidade, alinhados ao escopo específico da pesquisa. A inclusão de diversos tipos de documentos permite uma análise mais completa do panorama tecnológico e do impacto dos registros nos diferentes setores.

Em um cenário mais especializado, a avaliação de áreas específicas (Critério 03) assegura que o repositório abrange setores de interesse, como química, biotecnologia e farmacêutica, que apresentam particularidades distintas e demandam repositórios específicos que considerem essas características. Essa especialização viabiliza uma análise mais direcionada e também fornece dados com maior profundidade sobre áreas de interesse estratégico para o estudo.

De maneira geral, a avaliação da frequência de atualização e manutenção dos dados (Critério 04) é indispensável para evitar a utilização de informações desatualizadas ou de dados cuja atualização ocorra em intervalos prolongados, os quais não atendem às demandas da investigação. A adoção de repositórios com atualizações regulares proporciona uma análise mais precisa e contemporânea, assegurando que as conclusões da pesquisa se fundamentam nas informações mais recentes disponíveis.

De forma semelhante, uma interface de busca intuitiva (Critério 05), que oferece diversas opções de filtragem, facilita a formulação de consultas que possibilitam a localização e seleção de documentos de patentes relevantes para a pesquisa, otimizando assim o processamento dos dados. Esse aspecto é particularmente importante para a coleta de grandes conjuntos de dados, uma vez que possibilita a coleta incremental, por meio da aplicação de filtros que segmentam todo o conjunto disponível.

Avaliar os recursos de análise (Critério 06) disponíveis no repositório permite a realização de análises diretamente no repositório, eliminando a necessidade de transferência de dados para outras plataformas.

Considerar a disponibilidade de patentes em idiomas de interesse (Critério 07) é relevante para a condução eficiente de análises textuais, uma vez que permite levar em conta as particularidades de cada idioma sem a necessidade de traduções. Além disso, essa característica é relevante em estudos que envolvem mineração de texto ou análises textuais.

O custo de acesso (Critério 08) representa um fator significativo a ser considerado, uma vez que repositórios que exigem assinatura para acesso aos dados frequentemente disponibilizam recursos que não estão disponíveis em repositórios de acesso aberto. Contudo, esse aspecto deve ser analisado de maneira criteriosa. Ao avaliar a relação custo-benefício, o pesquisador pode garantir que o investimento no repositório seja proporcional ao valor dos dados oferecidos e compatível com os limites orçamentários da pesquisa.

Além disso, é vital considerar a relevância e a reputação do repositório (Critério 09), uma vez que esses fatores são essenciais para validar a credibilidade dos dados utilizados na pesquisa, levando em conta as experiências prévias de outros pesquisadores. A popularidade de um repositório entre a comunidade acadêmica serve como um indicativo de sua confiabilidade. Ao escolher repositórios que possuem avaliações positivas e ampla aceitação, aumenta-se a segurança de que os dados obtidos são confiáveis.

Na sequência, avaliar os recursos de exportação de dados (Critério 10) possibilita a manipulação e a análise externa das informações. Contudo, é crucial considerar a amplitude dos dados que podem ser exportados, uma vez que a quantidade de registros disponíveis para exportação pode variar de acordo com o repositório, assim como o formato do arquivo gerado para a exportação.

Por fim, a disponibilidade de APIs e Webservices (Critério 11) facilita a integração de processos computacionais voltados para a coleta automatizada de dados, otimizando a eficiência do processo de coleta e manutenção do conjunto de dados para análise. Essa funcionalidade é indispensável para pesquisas que demandam um grande volume de dados ou atualizações frequentes, pois economiza tempo e aumenta a eficiência operacional.

A aplicação desses critérios tem o objetivo de garantir que o repositório escolhido seja adequado às necessidades da pesquisa, e não atestar se o repositório possui ou não o recurso em questão. Cada critério aborda aspectos essenciais que atestam que o repositório selecionado atenda aos critérios desejados para a realização de uma investigação acadêmica. Ao adotar esses critérios de seleção, o pesquisador minimiza o risco de obter dados insuficientes ou inadequados.

Por conseguinte, com os critérios definidos, torna-se necessário elaborar uma matriz em que as colunas representam os repositórios potenciais a serem utilizados na pesquisa, e as linhas, os respectivos critérios. Na interseção entre o critério e o repositório, deve-se atribuir a pontuação de 1 (um) quando o repositório atender ao critério, 0 (zero) quando o critério for irrelevante para o estudo, e -1 (menos um) quando o repositório não atender ao critério estabelecido. Ao final, deve-se realizar a soma das pontuações obtidas, de modo que o repositório com a maior pontuação seja aquele que melhor se adequa ao escopo da pesquisa.

Com isso, estabelece-se um sistema de avaliação que permite classificar a relevância dos repositórios em relação ao escopo da pesquisa. Para esclarecer a aplicação da matriz proposta, considere-se um escopo de pesquisa em que o objetivo seja recuperar todas as patentes depositadas no Brasil entre os anos de 1900 a 2024 ([Tabela 3](#)).

Tabela 3 – Matriz de Triagem de Repositórios de Patentes

|           |                           | Repositórios |           |                |         |           |             |                          |                    |
|-----------|---------------------------|--------------|-----------|----------------|---------|-----------|-------------|--------------------------|--------------------|
|           |                           | INPI         | Espacenet | Google Patents | Latipat | The Lens  | Patentscope | Derwent Innovation Index | Orbit Intelligence |
| Critérios | Cobertura Geográfica      | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Amplitudes de Documentos  | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Áreas especializadas      | 0            | <b>0</b>  | 0              | 0       | <b>0</b>  | 0           | 0                        | 0                  |
|           | Atualização e Manutenção  | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Interface de Busca        | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Recursos de Análise       | -1           | <b>-1</b> | -1             | -1      | <b>-1</b> | -1          | -1                       | -1                 |
|           | Idioma                    | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Custo                     | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | -1                       | -1                 |
|           | Relevância do repositório | 1            | <b>1</b>  | 1              | 1       | <b>1</b>  | 1           | 1                        | 1                  |
|           | Exportação de dados       | -1           | <b>1</b>  | 1              | 1       | <b>1</b>  | -1          | 1                        | 1                  |
|           | API                       | -1           | <b>1</b>  | -1             | -1      | <b>1</b>  | -1          | 1                        | 1                  |
|           | Total                     | 4            | <b>8</b>  | 6              | 6       | <b>8</b>  | 4           | 6                        | 6                  |

Conforme apresentado, foram selecionados sete repositórios de abrangência internacional e um repositório regional mantido pelo INPI, considerando o escopo da pesquisa. Cada repositório foi avaliado em relação aos 11 critérios estabelecidos, validando-se sua adequação aos requisitos expostos por cada um deles. Diante desse processo, os repositórios que mais se alinharam aos critérios desejados foram o Espacenet e o Lens, ambos com uma pontuação final de 8.

Quando um repositório se destaca entre os demais, o processo de avaliação se encerra; entretanto, em casos de empate, torna-se pertinente reavaliar todos os critérios para determinar qual dos repositórios empatados melhor atende aos requisitos estabelecidos. Nesse caso, considerando o tamanho da base de dados, o Espacenet disponibiliza 149 milhões de documentos de patente, enquanto o Lens oferece 133 milhões. Assim, o Espacenet se mostra mais adequado ao critério “Amplitudes de Documentos” e, portanto, é o repositório que melhor se adequa ao escopo de pesquisa no exemplo supracitado.

A matriz de triagem de repositórios de patentes, apresentada na [Tabela 3](#), pode ser adaptada conforme as especificidades da pesquisa, permitindo a remoção de critérios irrelevantes ou a inclusão de novos critérios. Essa flexibilidade é especialmente útil na seleção de fontes de dados complementares, pois as características das fontes de dados complementares podem variar de acordo com as necessidades de cada investigação.

### 5.2.2 Catálogo de fonte de dados complementares

Considera-se como fontes de dados complementares aqueles conjuntos que não integram os dados bibliográficos das patentes, ou seja, que fornecem informações adicionais às características bibliográficas diretamente associadas às patentes. Assim, o presente estudo estabelece um catálogo inicial de fontes de dados complementares. Sob uma perspectiva nacional, destacam-se as seguintes fontes como excelentes opções:

**Base de Dados sobre Propriedade Intelectual para Fins Estatísticos – BADEPI:** No contexto das patentes depositadas no Brasil, o INPI, com o apoio da Organização Mundial da Propriedade Intelectual (OMPI), desenvolveu o projeto “Criação de Base de Dados sobre Propriedade Intelectual para Fins Estatísticos – BADEPI”<sup>2</sup>, com o objetivo de disponibilizar uma base de dados estruturada sobre propriedade industrial. Essa base reúne informações abrangentes sobre patentes depositadas no Brasil. A versão mais recente (2025), “BADEPI v8.3 Patentes”, compila dados bibliográficos relativos a diferentes modalidades de patente, como patente de invenção (PI), modelo de utilidade (MU) e certificado de adição (C), abrangendo o período de 1997 a 2020.

**Currículos Registrados na Plataforma Lattes:** Embora represente um conjunto restrito dos proponentes de patentes, os currículos na Plataforma Lattes são recursos valiosos,

<sup>2</sup> <<https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/estatisticas/base-de-dados-sobre-propriedade-intelectual-para-fins-estatisticos-2013-badepi>>

pois permitem o acesso a informações complementares sobre os proponentes de patentes que normalmente não estão disponíveis nos dados bibliográficos. A plataforma fornece detalhes como formação acadêmica, endereço profissional, área de pesquisa e atuação, nacionalidade, entre outros. Essas informações possibilitam uma caracterização mais completa dos indivíduos envolvidos na produção de patentes, algo que os dados bibliográficos, isoladamente, não permitem. A Plataforma Lattes disponibiliza livremente esses currículos online, documentando a trajetória profissional, acadêmica e as produções científicas de cada indivíduo.

Na próxima subseção são descritas todas as etapas do processo de aquisição dos dados.

### 5.2.3 Coleta de dados

Após a conclusão do processo de seleção dos repositórios de patentes e das fontes de dados complementares, torna-se imprescindível definir as estratégias de seleção e aquisição dos dados a serem utilizados. Essa etapa, no entanto, não é trivial, considerando a ampla diversidade de funcionalidades que cada repositório oferece para a seleção e *download* dos dados patentários. Essa heterogeneidade inviabiliza a adoção de abordagens genéricas, uma vez que os procedimentos ideais variam conforme as especificidades de cada repositório.

Nesse contexto, o primeiro repositório contemplado pelo Protocolo CTDP é o Espacenet. A escolha de iniciar pelo Espacenet se justifica por sua relevância consolidada entre os repositórios e pela abrangência de sua base de dados, que é a maior entre os repositórios considerados neste estudo.

Dito isto, ao investigar as especificações para coleta de dados disponíveis na página da Espacenet, identificou-se que a plataforma oferece um recurso de pesquisa que permite ao usuário inserir até 10 termos distintos, os quais são buscados em diversos campos da patente. Esse recurso, no entanto, foi projetado para uso manual, sem suporte para consultas automatizadas ou recuperação em lote. Quando tais operações são necessárias, recomenda-se o uso do serviço OPS (*Open Patent Services*).

O OPS é um serviço de *web* oferecido pelo EPO (*European Patent Office*), que permite o acesso aos dados do banco de patentes via serviços *web*, empregando a arquitetura RESTful. Utilizando os padrões XML (*eXtensible Markup Language*) e JSON (*JavaScript Object Notation*) para a formatação dos dados de resposta, o serviço viabiliza a parametrização e o desenvolvimento de processos automatizados para coleta de dados. No entanto, é imprescindível que o usuário realize um cadastro prévio<sup>3</sup>, para obter as credenciais de acesso ao serviço.

<sup>3</sup> <<https://www.epo.org/en/searching-for-patents/data/web-services/ops>>

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, 2023).

É importante destacar que, conforme mencionado anteriormente, 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, disponível em <https://www.epo.org/en/searching-for-patents/data/web-services/ops>. Com base nas informações presentes na documentação e em conformidade com os critérios de coleta definidos pelo repositório, o [Algoritmo 1](#) ilustra o processo de consulta incremental para a recuperação de patentes.

**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

```

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 representa 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. A partir desse momento, se 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, é verificado se o total de resultados *vTotalResultados* é maior que o valor final do intervalo de resultados *vRangeFinal*. Se for, será preciso atualizar os intervalos de pesquisa para a próxima página de resultados. Caso contrário, deve-se definir 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 dia. Todo esse tratamento é devido ao fato da 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 necessário 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 aos 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. O [Algoritmo 2](#) apresenta o pseudocódigo que descreve o 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

```

---

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.

Com a aplicação da estratégia de coleta de dados aqui apresentada, é atestada a viabilidade técnica e metodológica do processo de coleta de dados no repositório Espacenet.

Dando sequência ao processo de aquisição de dados, os procedimentos para a coleta de fontes complementares podem variar conforme a natureza e o escopo da fonte adotada. No caso específico de patentes depositadas no Brasil, a Base de Dados Sobre Propriedade Intelectual para Fins Estatísticos (BADEPI) fornece informações adicionais relevantes, especialmente no que tange aos proponentes das patentes. A versão 8.3 da base BADEPI é composta por sete arquivos no formato CSV ([Tabela 4](#)).

Tabela 4 – Tabelas BADEPI

| Arquivo             | Descrição                                                                             | Total de registros |
|---------------------|---------------------------------------------------------------------------------------|--------------------|
| PTN_CLASSIFICACOES  | Classificações recebidas pelas as patentes                                            | 1.153.723          |
| PTN_DEPOSITANTES    | Informações sobre os depositantes                                                     | 514.624            |
| PTN_DEPOSITANTES_V2 | Informações sobre os depositantes, apenas os dados do SINPI (sistema interno do INPI) | 549.820            |
| PTN_DEPOSITOS       | Informações sobre o depósitos das patentes                                            | 611.227            |
| PTN_DESPACHOS       | Informações sobre movimentações no processo de depósito da patente                    | 2.499.570          |
| PTN_INVENTORES      | Informações sobre os inventores                                                       | 1.217.003          |
| PTN_PCTS            | Informações do pedido internacional depositado através do sistema PCT                 | 245.401            |

O arquivo “PTN\_CLASSIFICACOES” inclui as classificações atribuídas às patentes; “PTN\_DEPOSITANTES” e “PTN\_DEPOSITANTES\_V2” contêm dados dos depositantes; “PTN\_DEPOSITOS” registra informações sobre os depósitos; “PTN\_DESPACHOS” detalha movimentações processuais; “PTN\_INVENTORES” apresenta informações sobre os inventores; e, por fim, “PTN\_PCTS” fornece dados sobre pedidos internacionais realizados pelo sistema PCT ([INPI, 2021c](#)). Os dados do BADEPI são de extrema importância, pois permitem a identificação detalhada dos envolvidos, incluindo nome, país, município, CEP e parte do CPF/CNPJ dos inventores e depositantes das patentes.

Outra fonte adicional de relevância no âmbito das patentes brasileiras é representada pelos currículos registrados na Plataforma Lattes, os quais contêm dados relacionados a patentes. A coleta e análise desses dados podem ser realizadas por meio do *LattesDataExplorer* (DIAS, 2016), bem como pela Plataforma BrCris<sup>4</sup>, um sistema agregador mantido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). O BrCris permite a recuperação, certificação e visualização de dados e informações de diversos atores envolvidos na pesquisa científica no Brasil e oferece funcionalidades para download dos resultados em formato CSV.

A próxima subseção descreve os procedimentos de tratamento de dados realizados para integrar os diferentes conjuntos coletados, visando garantir a compatibilidade entre as fontes de informação.

#### 5.2.4 Limpeza e transformação dos dados

Com a coleta dos dados concluída, faz-se necessário unificar os dados coletados e estabelecer relacionamentos entre eles. Definir uma padronização dos dados é fundamental para garantir a consistência e a comparabilidade entre os diferentes conjuntos de dados. Todos os dados coletados são armazenados em arquivos organizados em diretórios distintos. Embora esta organização viabilize o processo de recuperação e armazenamento, extrair informações que dependem de relacionamento e/ou cruzamento de dados, exige alto custo computacional, para recuperar as informações necessárias dos arquivos.

Uma abordagem eficaz para promover a uniformidade entre fontes de dados heterogêneas consiste em definir um ou mais esquemas que representem uma visão abrangente de todos os conjuntos de dados coletados. Esse processo, conhecido como integração de esquemas, é caracterizado pela resolução de conflitos semânticos, descritivos e estruturais entre os esquemas das diversas fontes de dados a serem integradas. Nesse contexto, um esquema estabelece a estrutura dos dados, abrangendo aspectos como os nomes de tabelas ou entidades, campos, tipos de dados e seus respectivos relacionamentos.

Assim, com o objetivo de integrar os diferentes conjuntos de dados, os dados são transformados em entidades e armazenados em um banco de dados relacional. Para isso, é adotada a especificação de um esquema integrado fundamentado no Modelo Entidade-Relacionamento, introduzido por Peter Pin-Shan Chen em 1976 (DATE, 2003).

O esquema proposto neste estudo estabelece uma estrutura lógica que define a organização e a forma de armazenamento dos dados coletados. Ele compreende a descrição formal das tabelas (entidades), suas colunas (atributos), tipos de dados, restrições, relacionamentos entre as entidades. Desta forma, o esquema serve como um conjunto de regras que especifica como os dados são organizados e como as relações entre diferentes

---

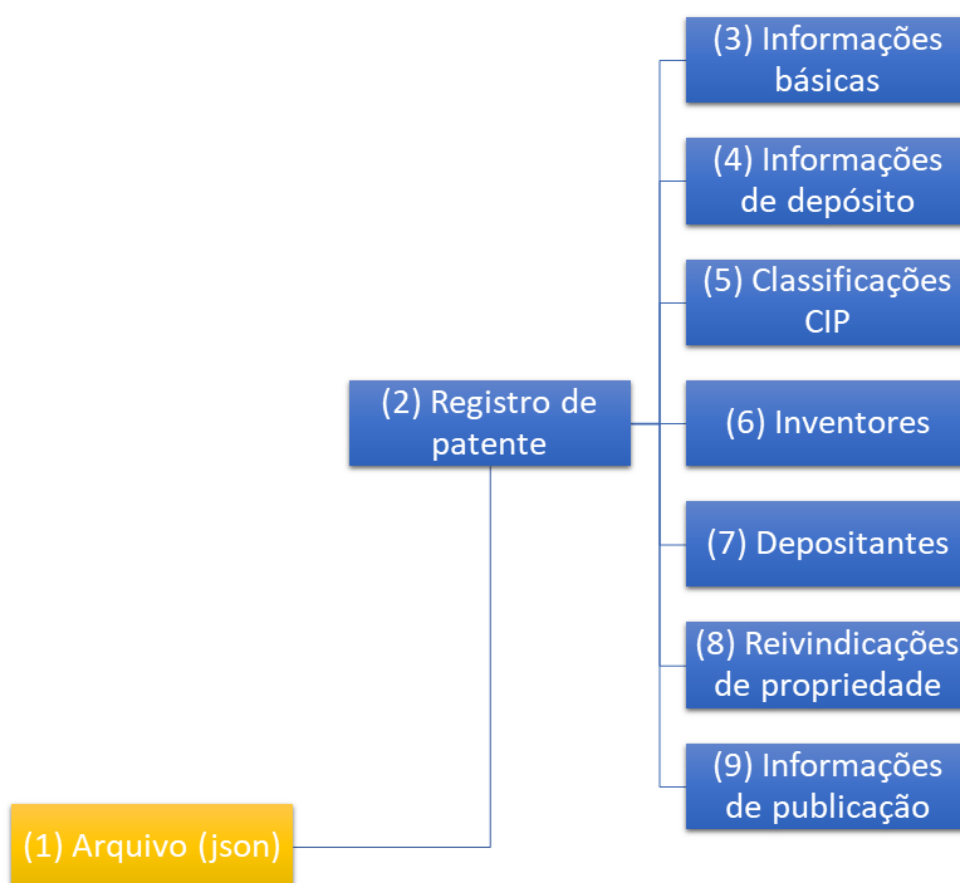
<sup>4</sup> <<https://brcris.ibict.br/>>

conjuntos de dados são estabelecidas.

Nesse sentido, uma entidade é considerada uma estrutura de dados que representa a essência de um tipo específico de informação. Por exemplo, a entidade "Patente" é uma estrutura de dados que agrupa as informações pertinentes a uma patente, constituindo assim um registro de patente no banco de dados. Em síntese, uma entidade é formada por atributos que a caracterizam, e esses atributos podem ser utilizados como critérios de seleção, bem como na composição das informações a serem extraídas do banco de dados.

Para a especificação das entidades que compõem o esquema, os conjuntos de dados coletados foram analisados com o intuito de identificar a estrutura atual dos mesmos. No que diz respeito ao conjunto de dados obtido na Espacenet, as patentes são disponibilizadas em arquivos no formato .json, sendo que cada arquivo contém informações sobre uma ou mais patentes. A [Figura 16](#) ilustra a organização do conteúdo de cada arquivo.

Figura 16 – Representação do arquivo de patentes no formato Espacenet



Fonte: Elaboração do autor

O item (1) ilustra o arquivo gerado com os dados obtidos ao consultar uma patente na Espacenet; já o item (2) representa um registro de patente; no item (3) são as informações básicas de identificação da patente, tais como título, resumo e código de identificação da

família a qual a patente pertence; o item (4) apresenta as informações de depósito da patente, como a data e número de depósito; no item (5) são as classificações recebidas pela a patente; no item (6), a listagem dos nomes dos inventores; o item (7) contém a listagem dos nomes dos depositantes; no item (8) informações das reivindicações de propriedade; e, por fim, no item (9) as informações de publicação da patente, como data de publicação e número de identificação da patente. Na documentação<sup>5</sup> disponibilizada pela Espacenet é possível obter mais informações sobre a estrutura dos arquivos de patente.

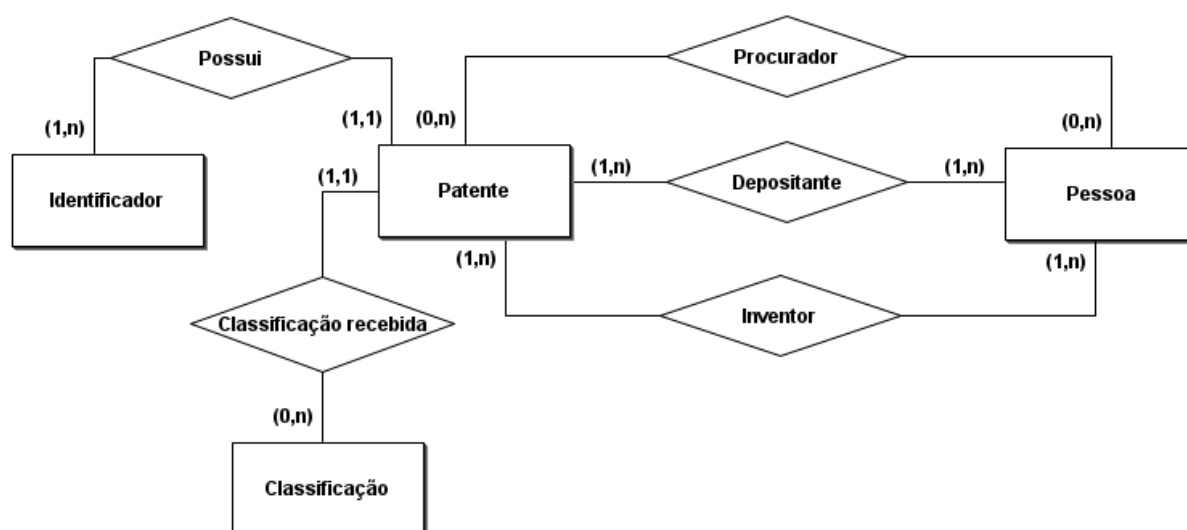
Assim como o conjunto de dados principal, os conjuntos de dados complementares coletados também devem ser analisados para identificar seu esquema atual de organização.

Após a compreensão do esquema adotado em cada conjunto de dados coletados, o próximo passo é mapear as entidades que farão parte do esquema único, ou seja, a construção de um único esquema para armazenar todos os dados coletados. O processo de mapeamento consiste em identificar as entidades que integrarão o esquema único, identificando entidades semelhantes em cada esquema e consolidando-as em uma única entidade que agregue todas as informações dos diferentes conjuntos. Esse mapeamento estabelece, portanto, um esquema integrado que abrange todos os esquemas individuais.

Diante do exposto, fundamentado nos insumos obtidos mediante a investigação dos esquemas adotado em cada conjunto de dados e posterior mapeamento, foi estabelecido um esquema único composto por duas categorias de entidades: as essenciais, derivadas dos dados coletados no repositório principal de patentes (neste caso, o repositório Espacenet), e as complementares, provenientes das fontes de dados complementares (INPI e Plataforma Lattes). Na versão inicial do Protocolo CTDP, o esquema integrado foi modelado a partir dos dados recuperados nas plataformas Espacenet, INPI e Lattes. O Modelo conceitual dos dados é apresentado pela [Figura 17](#).

<sup>5</sup> <<https://www.epo.org/en/searching-for-patents/data/web-services/ops>>

Figura 17 – Modelo conceitual da entidade definidas no esquema do protocolo CTD



A imagem mostra um diagrama de entidade-relacionamento (ER), que representa a modelagem conceitual dos relacionamentos entre entidades no contexto de patentes. As entidades apresentadas foram definidas a partir da fonte de dados principal, omitindo as entidades derivadas das fontes complementares para proporcionar uma visualização mais clara e objetiva. A entidade "Patente" é a entidade central do diagrama, detalhada no [Quadro 9](#). Ela estabelece diversas relações com outras entidades, como "Identificador", "Classificação" e "Pessoa".

Quadro 9 – Entidade Patente

| Atributos                  | Tipo                | Descrição                                                 |
|----------------------------|---------------------|-----------------------------------------------------------|
| ID                         | Numérico sequencial | Chave de identificação da entidade                        |
| Numero do documento        | Texto               | Número de depósito da patente                             |
| País                       | Texto               | Código do país de depósito da patente                     |
| Código de publicação       | Texto               | Código da versão de publicação da patente                 |
| Ultima publicação          | Sim/Não             | É a ultima versão de publicação da patente?               |
| Data de depósito           | Data                | Data de depósito da patente                               |
| Data de publicação         | Data                | Data de publicação da patente                             |
| Data de concessão          | Data                | Data de concessão da patente                              |
| Primeira Classificação CIP | Texto               | Código da primeira classificação da patente               |
| Título                     | Texto               | Título da patente                                         |
| Resumo                     | Texto               | Resumo da patente                                         |
| Identificação da família   | Texto               | Código de identificação da família da patente (Espacenet) |

A entidade "Patente" é composta por doze atributos que a caracterizam. Esses atributos foram definidos com base nos dados bibliográficos da patente, em conformidade com a Norma ST.9, exceto pelo campo ID, que representa um número sequencial destinado à identificação única de cada registro. A entidade "Patente" agrega os dados de "Patente",

"Depósito" e "Publicação" provenientes do conjunto de dados coletados na Espacenet, além de consolidar as informações relativas a "Depósito" e "Depósito PCT", extraídas do conjunto coletado no INPI.

A entidade Classificação ([Quadro 10](#)) indica as classificações associadas a uma patente.

Quadro 10 – Entidade Classificação

| Atributos     | Tipo                | Descrição                                         |
|---------------|---------------------|---------------------------------------------------|
| ID            | Numérico sequencial | Chave de identificação da entidade                |
| Classificação | Texto               | Classificação recebida pela patente               |
| Seção         | Texto               | Seção da classificação de acordo com a CIP        |
| Classe        | Texto               | Classe da classificação de acordo com a CIP       |
| Subclasse     | Texto               | Subclasse da classificação de acordo com a CIP    |
| Grupo         | Texto               | Grupo da classificação de acordo com a CIP        |
| Subgrupo      | Texto               | Subgrupo da classificação de acordo com a CIP     |
| Ordem         | Númerico            | Ordem de atribuição das classificações da patente |
| ID da patente | Númerico            | ID da patente que recebeu a classificação         |

Considerando que uma patente pode receber múltiplas classificações, tornou-se necessário definir uma entidade específica para representá-las. A entidade "Classificação" é composta por nove atributos, incluindo o valor da classificação e os valores de seus elementos constitutivos. O atributo "ordem" define a prioridade das classificações atribuídas à patente, enquanto o atributo "ID da patente" é utilizado para estabelecer a relação entre a patente e suas respectivas classificações.

A entidade "Pessoa" ([Quadro 11](#)) representa indivíduos que podem ter algum papel em relação à patente, como depositantes, procuradores ou inventores.

Quadro 11 – Entidade Pessoa

| Atributos       | Tipo                | Descrição                                  |
|-----------------|---------------------|--------------------------------------------|
| ID              | Numérico sequencial | Chave de identificação da entidade         |
| Nome            | Texto               | Nome do indivíduo                          |
| Pessoa Jurídica | SIM/NÃO             | Seção da classificação de acordo com a CIP |
| Nacionalidade   | Texto               | Nacionalidade do indivíduo                 |

Fonte: Elaboração do autor

A entidade "Pessoa" tem como objetivo centralizar todos os indivíduos envolvidos no processo de depósito, sejam eles pessoas físicas ou jurídicas, visando preparar um ambiente que possibilite a identificação única de cada participante. Para representar os relacionamentos entre as pessoas e as patentes, foram estabelecidas as entidades "Inventor", "Depositante" e "Procurador" conforme o detalhamento apresentado no [Quadro 12](#).

Quadro 12 – Entidades Depositante, Inventor e Procurador

| Entidade Depositante |                     |                                             |
|----------------------|---------------------|---------------------------------------------|
| Atributos            | Tipo                | Descrição                                   |
| Sequência            | Numérico sequencial | Ordem de participação informada no processo |
| ID da Pessoa         | Numérico            | ID do registro na Entidade Pessoa           |
| ID da Patente        | Numérico            | ID do registro na Entidade Patente          |
| Entidade Inventor    |                     |                                             |
| Sequência            | Numérico sequencial | Ordem de participação informada no processo |
| ID da Pessoa         | Numérico            | ID do registro na Entidade Pessoa           |
| ID da Patente        | Numérico            | ID do registro na Entidade Patente          |
| Entidade Procurador  |                     |                                             |
| Sequência            | Numérico sequencial | Ordem de participação informada no processo |
| ID da Pessoa         | Numérico            | ID do registro na Entidade Pessoa           |
| ID da Patente        | Numérico            | ID do registro na Entidade Patente          |

A entidade "Depositante" relaciona uma patente a seu(s) depositante(s). Uma pessoa pode ser depositante de uma ou mais patentes ((1,n)), enquanto uma patente deve ter ao menos um depositante (1,1). A entidade "Inventor" representa a relação entre uma patente e seu(s) inventor(es). Uma patente deve ter ao menos um inventor ((1,n)), e uma pessoa pode ser inventora de uma ou mais patentes. E, por fim, a entidade "Procurador" relaciona uma patente com o procurador responsável (caso aplicável). Aqui, uma pessoa pode ser procuradora de uma ou mais patentes, e uma patente pode ou não ter procurador, conforme indicado pela notação (0,n).

A entidade "Identificador" representa todas as formas de identificar uma patente, ou seja, todos os números de depósito e publicação atribuídos à patente durante todo o processo de depósito e publicação da patente. A Espacenet disponibiliza o número de publicação e números de depósito de uma patente que, em determinados casos, possuem valores diferentes, ou seja, uma mesma patente pode ter o número de depósito diferente do número de publicação. Desta forma, o objetivo desta entidade é armazenar todas as formas de identificar uma patente (Quadro 13).

Quadro 13 – Entidades Identificador

| Atributos               | Tipo     | Descrição                                            |
|-------------------------|----------|------------------------------------------------------|
| Número de identificação | Texto    | Identificação da patente (nº Depósito ou Publicação) |
| Origem do dado          | Texto    | A origem da informação                               |
| ID da Patente           | Numérico | ID do registro na Entidade Patente                   |

O protocolo proposto neste estudo, recomenda-se que as entidades essenciais já definidas não sejam alteradas; ao invés disso, deve-se criar vínculos por meio da inclusão de novas entidades. Para ilustrar, os dados complementares provenientes da Plataforma

Lattes podem ser modelados em duas entidades, sendo a primeira a entidade “Indivíduo Lattes” detalhada no [Quadro 14](#).

Quadro 14 – Entidade Indivíduo Lattes

| Atributos           | Tipo                | Descrição                                          |
|---------------------|---------------------|----------------------------------------------------|
| ID                  | Numérico sequencial | Chave de identificação da entidade                 |
| ID Lattes           | Texto               | Número de identificação do currículo na plataforma |
| Nome                | Texto               | Nome do proprietário do currículo                  |
| Data de Atualização | Data                | Data da ultima atualização do currículo            |
| ORCID               | Texto               | ORCID do individuo                                 |
| Pais                | Texto               | Pais de residência informado no currículo          |
| Estado              | Texto               | Estado de residência informado no currículo        |
| Cidade              | Texto               | Cidade de residência informado no currículo        |
| Area de Atuação     | Texto               | Primeira área de atuação informada                 |
| Maior Titulo        | Texto               | Maior titulação informada                          |
| ID da Pessoa        | Numérico            | ID do registro na Entidade Pessoa                  |

A entidade contém atributos extraídos dos currículos coletados na Plataforma Lattes, além do atributo “ID da Pessoa”, que permite estabelecer o vínculo entre o indivíduo mencionado no currículo e aquele participante do processo de depósito da patente. A segunda entidade, denominada “Patente Lattes”, é apresentada no [Quadro 15](#).

Quadro 15 – Entidade Patente Lattes

| Atributos               | Tipo                | Descrição                                        |
|-------------------------|---------------------|--------------------------------------------------|
| ID                      | Numérico sequencial | Chave de identificação da entidade               |
| Ano                     | Numérico            | Ano de depósito da Patente                       |
| Pais                    | Texto               | País de depósito                                 |
| Número                  | Texto               | Número de depósito da patente                    |
| Título                  | Texto               | Título da patente                                |
| Data de concessão       | Data                | Data de concessão da patente                     |
| Data de depósito        | Data                | Data de depósito da patente                      |
| Instituição de registro | Texto               | Instituição de registro da patente               |
| Depositante             | Texto               | Nome do depositante                              |
| Titular                 | Texto               | Nome do titular da patente                       |
| ID da Patente           | Numérico            | ID de vínculo com a Entidade Patente             |
| ID do Currículo Lattes  | Numérico            | ID do currículo de vínculo com a área de atuação |

A entidade “Patente Lattes” inclui atributos destinados a estabelecer vínculos entre as entidades, promovendo uma integração sem modificações na estrutura da entidade “Patente”. O atributo “ID da Patente” permite a conexão entre a patente presente na entidade “Patente Lattes” e a entidade “Patente”, agregando novas informações. Além disso, o atributo “ID do Currículo Lattes” cria uma associação com o currículo do proprietário da patente registrada, consolidando as relações entre os dados.

Concluído o processo de definição do esquema único, isto é, a modelagem conceitual dos dados, este pode servir como base para a construção do esquema lógico e a subsequente criação do banco de dados relacional. Ressalta-se que o protocolo CTDP não tem como objetivo apontar um sistema gerenciador de banco de dados (SGBD) específico, ele apenas define a necessidade de ser um banco de dados relacional, sem vinculação a uma tecnologia particular.

Após a definição do SGBD e a construção do banco de dados relacional, é necessário realizar o processo de carga dos dados. Esse processo consiste em percorrer o conjunto de dados coletados, transformá-los em entidades, validar essas entidades e armazená-las em suas respectivas tabelas. Para determinar se um registro de um tipo específico de entidade é válido, é essencial avaliar se ele possui os atributos mínimos necessários para caracterizar a entidade. Além de verificar a presença dos atributos, é fundamental analisar se o valor de cada atributo está no formato compatível com o esperado. Os critérios estabelecidos para avaliar a validade das entidades estão apresentados no [Quadro 16](#).

Quadro 16 – Critérios de validação

| Entidade               | Atributo                   | Exigências                                 |
|------------------------|----------------------------|--------------------------------------------|
| Patente                | Numero do documento        | Texto / Obrigatório                        |
|                        | País                       | Texto / Obrigatório / 2 Caracteres         |
|                        | Data de depósito           | Data / Formato "dd/mm/aaaa"                |
|                        | Data de publicação         |                                            |
|                        | Data de concessão          |                                            |
|                        | Primeira classificação CIP | Texto / Obrigatório                        |
|                        | Título                     |                                            |
| Entidade Classificação | Classificação              | Código de classificação válido formato CIP |
| Entidade Pessoa        | Nome                       | Texto / Obrigatório                        |
| Entidade Depositante   | Sequência                  | Numérico / Obrigatório                     |
|                        | ID da Pessoa               |                                            |
|                        | ID da Patente              |                                            |
| Entidade Inventor      | Sequência                  | Numérico / Obrigatório                     |
|                        | ID da Pessoa               |                                            |
|                        | ID da Patente              |                                            |
| Entidade Procurador    | Sequência                  | Numérico / Obrigatório                     |
|                        | ID da Pessoa               |                                            |
|                        | ID da Patente              |                                            |
| Entidade Identificador | Número de identificação    | Texto / Obrigatório                        |
|                        | ID da Patente              | Numérico / Obrigatório                     |

O quadro apresenta os atributos obrigatórios para cada tipo de entidade, assim como a definição dos formatos esperados. O processo de validação das entidades, portanto,

consiste em verificar se cada registro, resultante da transformação dos dados, está em conformidade com os critérios apresentados no quadro. Para os dados complementares, também é necessário estabelecer critérios de validação, que podem variar conforme o tipo de dado complementar, mas é fundamental que elenquem atributos que caracterizam o tipo de dado a ser carregado. Assim, apenas as entidades consideradas válidas devem ser armazenadas no banco de dados.

Após a inclusão das entidades consideradas válidas no banco de dados, deve ser realizado o processo de deduplicação dos dados, termo este que se refere ao processo de identificar e eliminar registros duplicados em conjuntos de dados. O processo de deduplicação é caracterizado por quatro etapas distintas, a saber: (1) divisão de dados; (2) impressões digitais; (3) comparação; e, por fim, (4) eliminação de duplicatas.

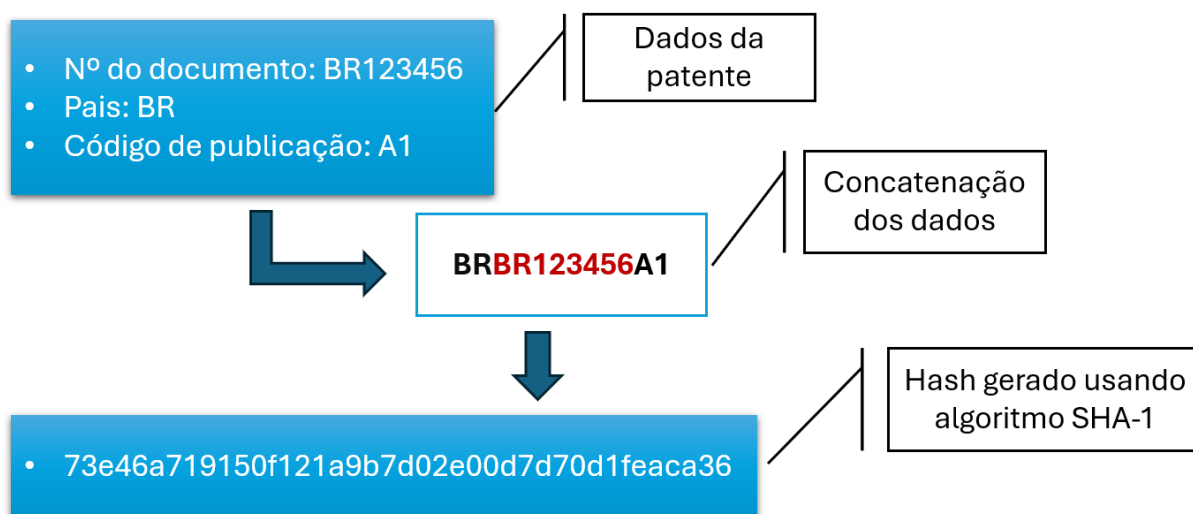
Na etapa inicial de deduplicação, realiza-se a divisão dos dados, agrupando-os por tipo de entidade, o que gera conjuntos menores de toda a base de dados. Em seguida, na segunda etapa, são selecionados atributos específicos de cada entidade para a construção das impressões digitais, ou identificadores únicos, conhecidos como *hash*. *Hash* são funções criptográficas, como MD5 (Message Digest 5) e SHA-1 (Secure Hash Algorithm 1). São algoritmos matemáticos determinísticos projetados para transformar dados de tamanho arbitrário em uma sequência de comprimento fixo, denominada “*hash*”. Esse processo busca garantir propriedades como integridade e resistência a colisões, ou seja, encontrar duas entradas diferentes com o mesmo hash. E quaisquer alterações nos dados de entrada produzem uma saída significativamente distinta.

O [Quadro 17](#) apresenta os atributos escolhidos para a criação do *hash* de cada entidade essencial, assegurando a unicidade dos registros e facilitando o processo de deduplicação.

Quadro 17 – Critérios de deduplicação

| Nº | Entidade      | Atributos                                       |
|----|---------------|-------------------------------------------------|
| 01 | Patentes      | País; Número do documento; Código de publicação |
| 02 | Classificação | Ordem; Classificação;                           |
| 03 | Pessoa        | Nome                                            |

A [Figura 18](#) apresenta um exemplo de construção de um *hash* de uma patente.

Figura 18 – Exemplo de construção de um *hash* de uma patente

Fonte: Elaboração do autor

Conforme apresentado, os dados listados no [Quadro 17](#), para a construção do *hash*, são concatenados na ordem definida. No caso da patente, conforme ilustrado na [Figura 18](#), primeiro aparece a sigla do país, seguida pelo número do documento e pelo código de publicação. Concluída a construção do *hash*, passa-se à terceira etapa, onde os *hashes* são comparados entre si para identificar chaves idênticas. Finalmente, na quarta etapa, ao identificar duplicatas, apenas a cópia mais completa é preservada, ou seja, aquela com o maior número de atributos preenchidos é mantida, enquanto as demais são eliminadas do banco de dados.

Conforme as estratégias discutidas nesta subseção, o Protocolo CTDTP estabelece um conjunto estruturado de procedimentos que viabiliza a integração de múltiplos conjuntos de dados.

Na subseção seguinte, são apresentados os métodos adotados para consultar as informações no banco de dados resultante dessa integração.

### 5.2.5 Consulta e seleção de dados

No módulo de exportação, são estabelecidas estratégias para recuperação dos dados. A adoção de um banco de dados relacional possibilita o uso de SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada, em português) para consultar os dados armazenados. A SQL é uma linguagem de programação especializada, projetada para gerenciar e manipular informações em bancos de dados relacionais, permitindo realizar operações como inserção (*INSERT*), atualização (*UPDATE*), exclusão (*DELETE*) e consulta (*SELECT*). Embora o uso da SQL exija conhecimentos técnicos, essa linguagem oferece maior flexibilidade para a recuperação de informações, pois permite o uso de qualquer

atributo de quaisquer entidades definidas como critério de busca, sem limitar as consultas aos campos disponibilizados nas interfaces dos repositórios de patentes disponíveis na internet.

Consequentemente, a adoção de um banco de dados relacional facilita a integração com ferramentas compatíveis, como o Microsoft Power BI<sup>6</sup>, um serviço de análise de dados. Além disso, permite a extração de subconjuntos de dados para armazenamento em arquivos tabulados, possibilitando o uso de ferramentas de análise visual, como o Gephi<sup>7</sup> e o VOSviewer<sup>8</sup>.

A qualificação do Protocolo de Coleta e Tratamento de Dados Patentários (Protocolo CTDTP) foi obtida por meio das estratégias metodológicas apresentadas. O protocolo se vale de técnicas e ferramentas amplamente consolidadas no mercado tecnológico e acadêmico, sendo justificado por sua capacidade de fornecer soluções eficazes e acessíveis para o tratamento de dados bibliográficos de patentes, aspecto de grande importância no campo da propriedade intelectual.

Ainda que não represente funcionalidades inéditas, o uso integrado e orientado dessas ferramentas e técnicas é importante, pois viabiliza a sistematização e formalização de processos que antes eram realizados de forma dispersa ou não padronizada.

Na próxima seção, é apresentado o estudo de praticabilidade e eficácia do protocolo CTDTP.

### 5.3 Estudo de praticabilidade e eficácia

Esta seção destina-se a apresentar o estudo de praticabilidade e eficácia realizado com o protocolo proposto nesta tese. Tal estudo busca responder a duas questões fundamentais acerca do protocolo CTDTP:

1. **Praticabilidade:** É possível executar o protocolo com os recursos disponíveis e dentro das restrições impostas?
2. **Eficácia:** Se o protocolo for executado, ele alcançará os resultados esperados?

Para a realização do estudo de viabilidade e eficácia do protocolo CTDTP, desenvolveu-se uma ferramenta que aplica integralmente os princípios estabelecidos por este protocolo. A subseção a seguir apresenta em detalhes os aspectos técnicos envolvidos no desenvolvimento da ferramenta, incluindo sua arquitetura, os processos de coleta, tratamento e armazenamento dos dados.

<sup>6</sup> <<https://www.microsoft.com/pt-br/power-platform/products/power-bi>>

<sup>7</sup> <<https://gephi.org/>>

<sup>8</sup> <<https://www.vosviewer.com/>>

### 5.3.1 Pipeline para coleta e tratamento de dados de patentes

Para o estudo de praticabilidade e eficácia, foi desenvolvida nesta tese a ferramenta denominada "Pytente". A escolha do nome "Pytente" reflete uma combinação estratégica que destaca tanto a base tecnológica quanto o propósito da ferramenta: a coleta e análise de dados bibliográficos de patentes. O nome resulta da junção das palavras "Python" e "Patente", evidenciando a linguagem de programação utilizada no desenvolvimento e a natureza das informações manipuladas. Dessa forma, o nome comunica diretamente a função e o objetivo da ferramenta, facilitando o reconhecimento e a associação imediata ao seu propósito principal.

A escolha da linguagem de programação Python para o desenvolvimento da ferramenta Pytente 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.

Para a construção do banco de dados local, optou-se pelo uso do 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, 2023](#)).

Para validar a viabilidade e eficácia do protocolo CTDp, sua aplicação foi realizada por meio da ferramenta Pytente, com o objetivo de coletar o maior número possível de patentes depositadas no Brasil no período compreendido entre 1900 e 2024. Essa iniciativa teve como propósito a obtenção de um extenso conjunto de dados que, de outra forma, não estaria prontamente acessível apenas por meio de repositórios de patentes de acesso aberto. Com os dados obtidos, pretende-se elaborar um panorama que caracterize a produção técnica brasileira, utilizando os documentos de patentes como principal fonte de análise.

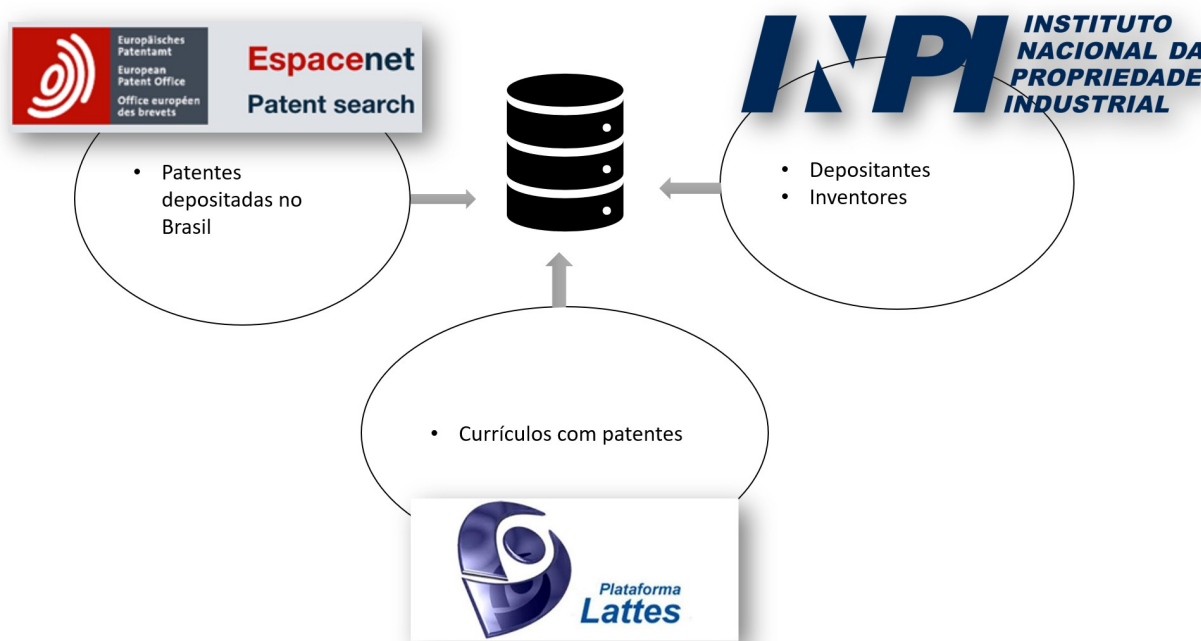
Dito isto, na primeira etapa de execução do protocolo CTDp, são definidas as fontes de dados. Para a construção da base de dados local foram agregados dados de três diferentes fontes, Espacenet<sup>9</sup> como fonte de dados principal e o INPI<sup>10</sup> e Plataforma Lattes<sup>11</sup> ([Figura 19](#)) como fonte de dados complementares.

<sup>9</sup> <<https://worldwide.espacenet.com/>>

<sup>10</sup> <<https://busca.inpi.gov.br/pePI/>>

<sup>11</sup> <<https://www.lattes.cnpq.br/>>

Figura 19 – Repositório de dados selecionados para a construção da base de dados local

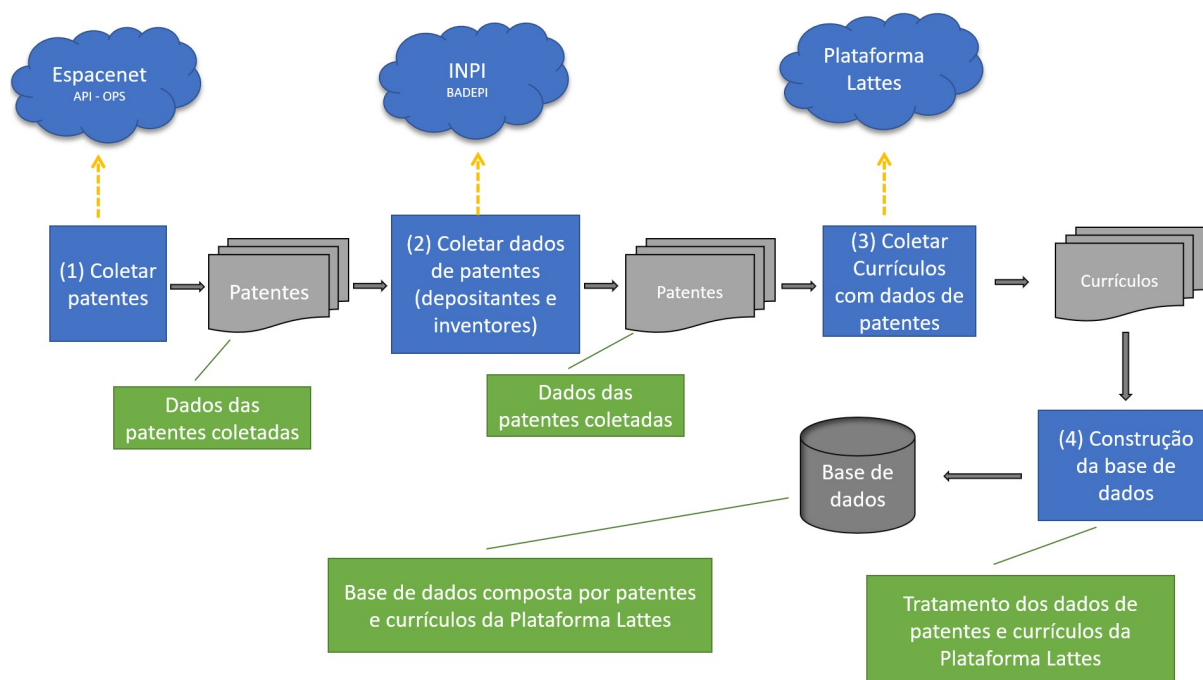


Fonte: Elaboração do autor

A seleção da Espacenet foi guiada pela Matriz de Triagem de Repositórios de Patentes proposta pelo protocolo CTDTP. Já a escolha do INPI e da Plataforma Lattes como fontes de dados complementares revela-se estratégica, pois, no escopo do Estudo de Praticabilidade e Eficácia, esses repositórios são essenciais para integrar a coleta de patentes depositadas no Brasil.

O processo de coleta dos dados foi dividido em quatro etapas distintas, (1) coleta dos dados de patentes no repositório da Espacenet, (2) coleta de patentes no INPI, (3) a coleta de dados dos currículos da Plataforma Lattes, que possuem referências a depósito de patentes, e, por fim, (4) a construção da base de dados local. A [Figura 20](#) apresenta a visão geral do esquema elaborado para o processo de coleta dos dados.

Figura 20 – Visão geral do processo de coleta de dados

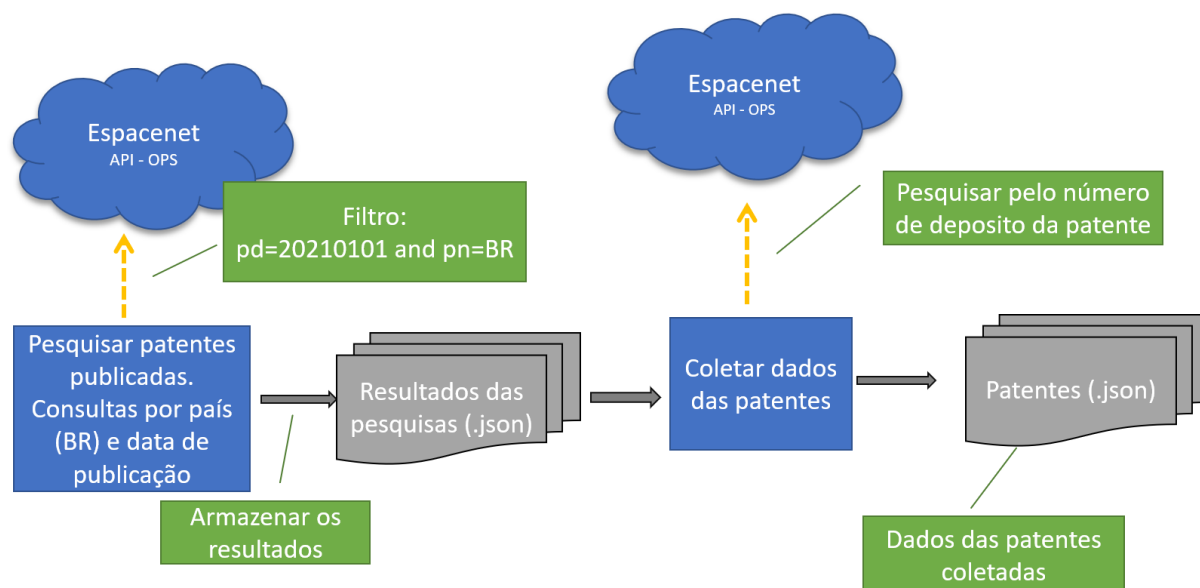


Fonte: Elaboração do autor

Conforme apresentado na figura, foi implementado o [Algoritmo 1](#), mencionado na [Subseção 5.2.3](#), utilizando a linguagem de programação Python. O algoritmo realiza requisições aos serviços de pesquisa de patentes da OPS, especificando a data de publicação e o país, efetuando uma requisição para cada dia dentro do período de janeiro de 1900 a novembro de 2024. Essa abordagem segue a estratégia definida pelo protocolo CTDTP para maximizar a recuperação de patentes.

Como resultado da execução do algoritmo, foram gerados diversos arquivos contendo os números de depósito das patentes que atenderam aos filtros estabelecidos. Em seguida, com a implementação do [Algoritmo 2](#), foi realizada uma requisição ao serviço “*publication*” da OPS para cada patente, usando como parâmetro o número de depósito, por exemplo, “BR 9104913A”. Esse procedimento teve como objetivo obter os dados bibliográficos das patentes obtidas mediante a execução do primeiro algoritmo. A [Figura 20](#) fornece uma visão detalhada do processo de coleta de patentes na Espacenet.

Figura 21 – Coleta de patentes na Espacenet



Fonte: Elaboração do autor

Conforme apresentado na figura, o fluxo do processo divide-se em duas fases principais: a busca inicial de patentes publicadas, baseada em critérios específicos, e a coleta detalhada dos dados das patentes identificadas, utilizando o número de depósito para consultas individuais. Na primeira etapa, representada pelo primeiro bloco azul, realiza-se uma busca por patentes publicadas no Brasil (código de país "BR"), a partir de uma data de publicação determinada. Os resultados dessas buscas são armazenados em arquivos `.json`, contendo as informações das patentes que atendem aos critérios estabelecidos.

Após identificar as patentes de interesse, inicia-se a segunda fase, em que são coletados dados bibliográficos de cada patente. Para isso, consulta-se novamente a API do Espacenet, agora usando o número de depósito de cada patente (sua identificação única). Os dados bibliográficos das patentes coletadas são organizados e armazenados em arquivos `.json`, proporcionando um repositório estruturado dos dados obtidos.

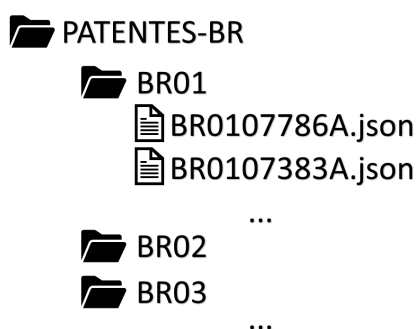
O código implementado permite manter a base de dados constantemente atualizada, realizando consultas de patentes por data de publicação para obter as patentes divulgadas no período especificado. Além disso, patentes já coletadas também são atualizadas, pois, ao serem modificadas, recebem uma nova data de publicação, o que possibilita identificá-las e recuperar seus dados atualizados.

O conjunto de patentes coletadas na Espacenet, conforme descrito anteriormente, é armazenado em arquivos no formato `.json`, o que facilita a manipulação dos dados. Todos os arquivos são organizados em um diretório principal denominado "PATENTES-BR". Cada arquivo recebe o nome correspondente ao número da patente que armazena, como, por

exemplo, BR0107786A.json.

Para otimizar a organização e o acesso a esses arquivos, foram criados subdiretórios baseados nos quatro primeiros caracteres do nome do arquivo. Assim, no exemplo mencionado, o arquivo BR0107786A.json é armazenado em um subdiretório nomeado BR01. A estrutura esquemática dessa organização encontra-se ilustrada na [Figura 22](#).

Figura 22 – Organização de arquivos de patente



Fonte: Elaboração do autor

A estrutura de diretórios apresentada na figura permite uma organização eficiente dos arquivos, otimizando tanto o armazenamento quanto a recuperação dos dados. A [Tabela 5](#) apresenta uma amostra dos diretórios criados e a quantidade de arquivos armazenados em cada um.

Tabela 5 – Listagem de pastas

| Diretório | Quantidade | Diretório | Quantidade | Diretório | Quantidade |
|-----------|------------|-----------|------------|-----------|------------|
| BR00      | 17.372     | BR60      | 1.760      | BR82      | 11.334     |
| BR01      | 17.090     | BR61      | 1.861      | BR83      | 10.894     |
| BR02      | 15.696     | BR62      | 1.730      | BR84      | 7.298      |
| BR03      | 17.227     | BR63      | 2.017      | BR85      | 7.258      |
| BR04      | 3          | BR64      | 2.149      | BR86      | 7.137      |
| BR05      | 1          | BR65      | 2.129      | BR87      | 7.812      |
| BR10      | 58.994     | BR66      | 2.584      | BR88      | 7.853      |
| BR11      | 186.469    | BR67      | 2.930      | BR89      | 7.791      |
| BR12      | 3.398      | BR68      | 2.810      | BR90      | 7.959      |
| BR13      | 618        | BR69      | 2.662      | BR91      | 7.222      |
| BR20      | 17.052     | BR70      | 2.872      | BR92      | 6.967      |
| BR21      | 255        | BR71      | 2.867      | BR93      | 7.673      |
| BR22      | 17         | BR72      | 2.117      | BR94      | 8.507      |
| BR30      | 5.922      | BR73      | 2.557      | BR95      | 10.527     |
| BR32      | 12         | BR74      | 5.646      | BR96      | 12.677     |
| BR53      | 1          | BR75      | 11.307     | BR97      | 14.992     |
| BR54      | 957        | BR76      | 11.479     | BR98      | 16.072     |
| BR55      | 1.479      | BR77      | 11.668     | BR99      | 17.578     |
| BR56      | 1.447      | BR78      | 11.553     | BRC1      | 1          |
| BR57      | 1.639      | BR79      | 12.221     | BRDI      | 850        |
| BR58      | 1.817      | BR80      | 12.247     | BRMU      | 24.146     |
| BR59      | 1.802      | BR81      | 12.322     | BRPI      | 153.431    |

Os arquivos coletados foram distribuídos em 66 diretórios. Nota-se na tabela que a sequência dos diretórios não é contínua, pois alguns ficaram vazios e foram omitidos.

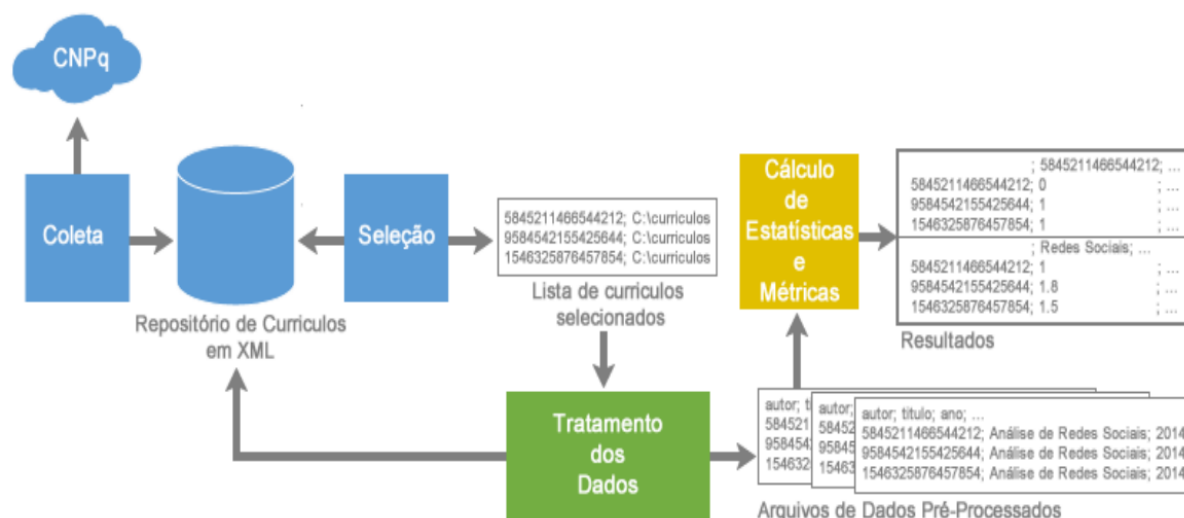
Quanto às fontes de dados complementares, a coleta junto ao INPI é caracterizada pela a coleta de dados de patentes a partir da base “BADEPI v8.3 Patentes”. A versão mais recente do BADEPI pode ser obtida na página do INPI<sup>12</sup>; contudo, utilizou-se neste estudo a versão 8.3.

Na sequência do processo de aquisição de dados, foi realizada a coleta de currículos registrados na Plataforma Lattes, que contém informações de patentes, como número de depósito ou título. O objetivo dessa etapa foi agregar dados sobre os indivíduos e validar essas informações em relação ao conjunto previamente coletado na Espacenet e no INPI. Essa validação é essencial, pois as informações na Plataforma Lattes são fornecidas pelo próprio usuário, exigindo, portanto, um processo de verificação para assegurar a confiabilidade e a certificação dos dados.

A coleta e seleção de dados curriculares dessa plataforma foram realizadas com o auxílio do *framework LattesDataXplorer* (DIAS, 2016), que agrega um conjunto de técnicas e métodos para coleta, seleção, tratamento e análise de dados (Figura 23).

<sup>12</sup> <<https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/estatisticas/base-de-dados-sobre-propriedade-intelectual-para-fins-estatisticos-2013-badepi>>

Figura 23 – Visão geral do LattesDataXplorer



Fonte: Dias (2016)

O *framework LattesDataXplorer* é composto por diversos módulos responsáveis por extrair e selecionar os dados curriculares da Plataforma Lattes. Neste estudo foram utilizados os módulos coleta e seleção. O processo de coleta é composto por 3 etapas:

1. **Extração de URLs:** responsável por obter os códigos únicos de todos os currículos cadastrados, possibilitando o acesso individual a cada currículo. Cada currículo na Plataforma Lattes possui um código único composto de oito dígitos.
2. **Extração de Identificadores e Data:** Acessar o cabeçalho de cada currículo para recuperar seu identificador único, a data de última atualização e o código do currículo, armazenando os dados localmente em um arquivo de identificação que contém código, identificador, data da última atualização no CNPq e data da atualização do currículo.
3. **Extração dos currículos:** Extrair e armazenar os currículos. Caso a data de atualização na Plataforma Lattes seja divergente da data de atualização do currículo armazenado localmente, o currículo local é substituído pelo mais recente. Os currículos coletados possuem o formato XML, pois essa versão contém seções e campos bem delimitados.

O extrator coleta os currículos e os armazena no formato XML em pastas identificadas de "00" a "99". A pasta selecionada e o nome do arquivo são definidos de acordo com o identificador único de 16 dígitos do currículo: os dois primeiros dígitos correspondem ao nome da pasta, enquanto os 14 dígitos restantes formam o nome do arquivo.

Com a conclusão da coleta de dados nos repositórios Espacenet e INPI, bem como nos currículos da Plataforma Lattes, torna-se necessário integrar esses diferentes conjuntos de dados. Para isso, é fundamental realizar o tratamento adequado das informações

coletadas. Esse processo de tratamento inclui etapas como a organização dos dados em um formato apropriado para análise, a limpeza para correção de possíveis erros e remoção de valores ausentes, além da transformação dos dados em entidades, e, posteriormente, o armazenamento em um banco de dados relacional.

Assim, o banco de dados local, fundamentado nas entidades essenciais propostas pelo protocolo CTDP e com a inclusão de novas entidades complementares, é composto pelas tabelas descritas no [Quadro 18](#). Para facilitar a visualização, a estrutura do banco de dados local é apresentada em seções: “Parte A” contém os dados de patentes obtidos na Espacenet; “Parte B” reúne dados complementares de patentes provenientes do INPI; e, por fim, “Parte C” agrega dados extraídos dos currículos da Plataforma Lattes.

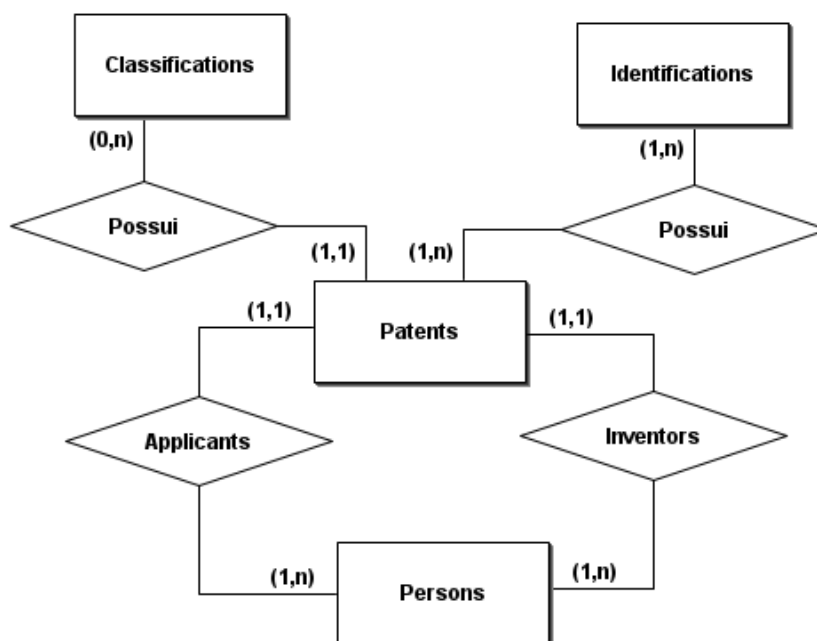
Quadro 18 – Descrição das tabelas que compõem banco de dados local

| Agrupamentos | Tabela                   | Descrição                                                                              |
|--------------|--------------------------|----------------------------------------------------------------------------------------|
| Grupo A      | PATENTS                  | Informações básicas de uma patente                                                     |
|              | PERSONS                  | Informações de identificação de toda as pessoas envolvidas (depositantes e inventores) |
|              | APPLICANTS               | Relacionamento do tipo depositante entre as tabelas “PATENTS” e “PERSONS”              |
|              | INVENTORS                | Relacionamento do tipo inventor entre as tabelas “PATENTS” e “PERSONS”                 |
|              | CLASSIFICATIONS          | Informações das classificações recebidas pelas patentes                                |
|              | IDENTIFICATIONS          | Armazena todas as formas de se identificar uma patente                                 |
|              | PERSON_CITATIONS         | Armazena as diferentes formas em que uma pessoa é referenciada                         |
| Grupo B      | INPI_PATENT              | Informações de identificação da patente no INPI                                        |
|              | INPI_INVENTOR            | Informações de identificação dos inventores das patentes                               |
|              | INPI_DEPOSITOR           | Informações de identificação dos depositantes das patentes                             |
| Grupo C      | LATTES_IDENTIFICATION    | Identificação do currículo                                                             |
|              | LATTES_PATENTS           | Patente informada nos currículos da Plataforma Lattes                                  |
|              | LATTES_PATENTS_AUTHORS   | Inventores das patentes informados nos currículos da Plataforma Lattes                 |
|              | LATTES_SOFTWARE          | Programas de computador com registro informados na Plataforma Lattes                   |
|              | LATTES_SOFTWARE_AUTHORS  | Autores dos programas de computador com registro                                       |
|              | LATTES_OCCUPATION_AREA   | Áreas de atuação informada nos currículos                                              |
|              | LATTES_FORMATION         | Formação dos proprietários dos currículos                                              |
|              | LATTES_AREA_OF_EXPERTISE | Áreas de atuação informadas nos currículos                                             |
|              | LATTES_ADDRESS           | Endereço profissional informado nos currículos                                         |

Fonte: Elaboração do autor

No “Grupo A”, os dados coletados na Espacenet foram organizados em sete tabelas, conforme descrito a seguir. A tabela “PATENTS” armazena as informações básicas de uma patente, incluindo identificação, família, título e resumo. Na tabela “PERSONS”, são registrados os nomes de todas as pessoas envolvidas no processo, seja como depositantes ou inventores. A tabela “APPLICANTS” estabelece o relacionamento de tipo “depositante” entre as tabelas “PATENTS” e “PERSONS”, enquanto a tabela “INVENTORS” representa o relacionamento de tipo “inventor” entre essas mesmas tabelas. A tabela “CLASSIFICATIONS” registra todas as classificações atribuídas às patentes. Na tabela “IDENTIFICATIONS”, são armazenadas as diferentes formas de identificação de uma patente, incluindo variações nos números de identificação, caso existam. Por fim, a tabela “PERSON\_CITATIONS” armazena as diferentes grafias de um mesmo nome, registrando as várias maneiras pelas quais uma pessoa é referenciada. A [Figura 24](#) apresenta o modelo conceitual do Grupo A no banco de dados.

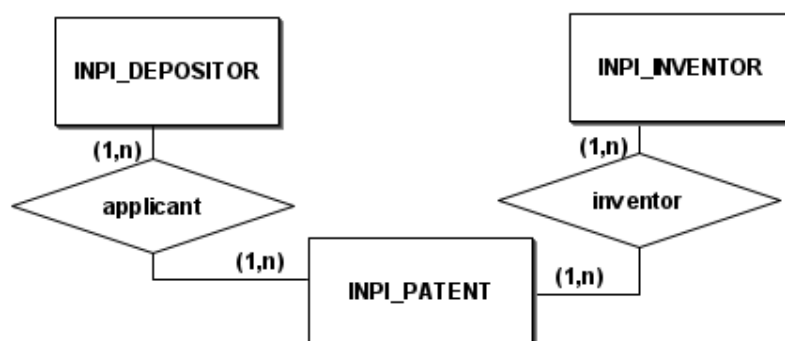
Figura 24 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo A



Fonte: Elaboração do autor

No "Grupo B", os dados coletados no INPI, referentes ao depósito de patentes, foram organizados em três tabelas. A tabela "INPI\_PATENT" armazena as informações de identificação de cada patente; a tabela "INPI\_INVENTOR" registra os dados dos inventores; e, por fim, a tabela "INPI\_DEPOSITOR" contém as informações dos depositantes. A [Figura 25](#) ilustra o Grupo B do modelo conceitual do banco de dados.

Figura 25 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo B

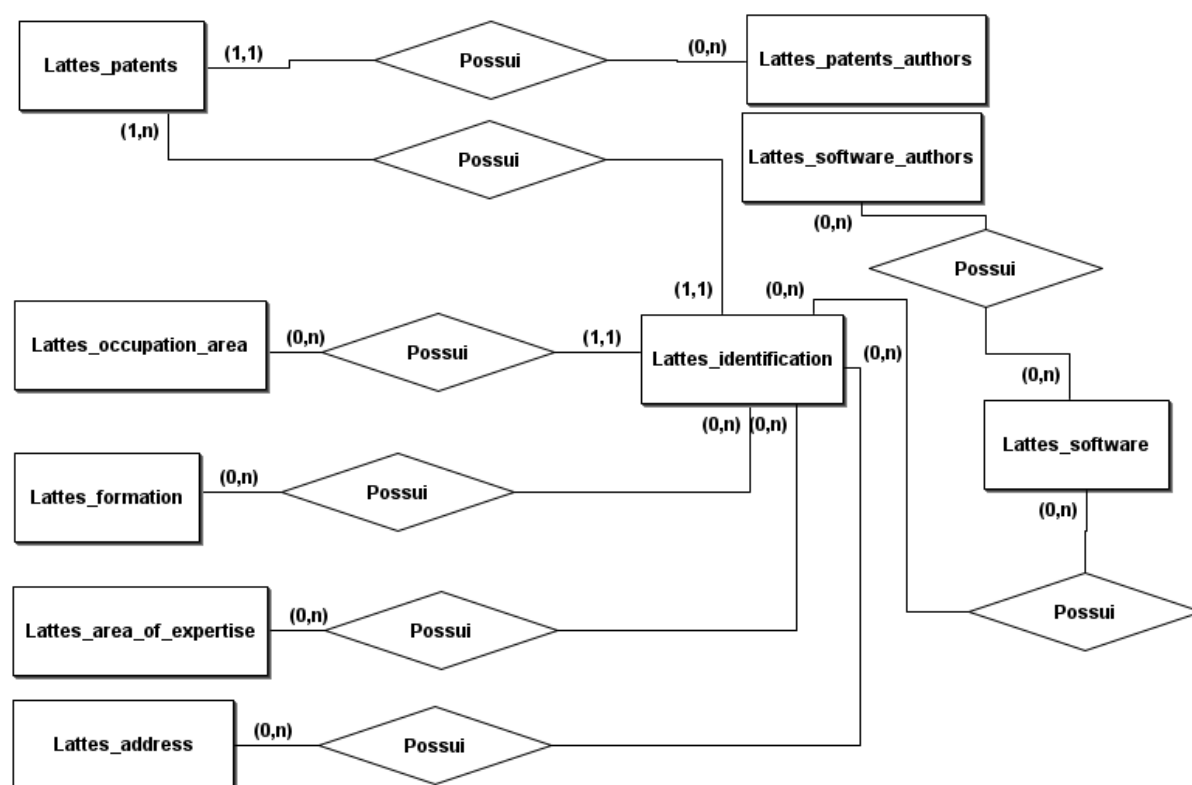


Fonte: Elaboração do autor

Por fim, os dados coletados na Plataforma Lattes foram organizados em nove tabelas. A tabela "LATTES\_IDENTIFICATION" contém as informações de identificação do

currículo. A tabela “LATTES\_PATENTS” armazena os dados sobre as patentes mencionadas nos currículos, e está relacionada com a tabela “LATTES\_PATENTS\_AUTHORS”, onde são registrados os inventores dessas patentes. A tabela “LATTES\_SOFTWARE” reúne as informações sobre programas de computador com registro e mantém relacionamento com a tabela “LATTES\_SOFTWARE\_AUTHORS”, que contém os dados dos autores desses programas. A tabela “LATTES\_OCCUPATION\_AREA” armazena as áreas de atuação informadas nos currículos, enquanto a tabela “LATTES\_FORMATION” guarda as informações sobre a formação acadêmica dos currículos. A tabela “LATTES\_AREA\_OF\_EXPERTISE” registra as áreas de especialização dos indivíduos, e, por fim, a tabela “LATTES\_ADDRESS” contém os endereços profissionais fornecidos. A [Figura 26](#) apresenta o Grupo C do modelo conceitual do banco de dados.

Figura 26 – Modelo conceitual do banco de dados das tabelas que compõem o Grupo C



Fonte: Elaboração do autor

Após a definição da estrutura do banco de dados, tornou-se necessário percorrer cada registro coletado e armazená-lo adequadamente. Para tanto, foi estabelecido um conjunto de critérios para determinar a validade de cada entrada de dados, ou seja, para definir o mínimo de informações que uma entrada deve conter para ser considerada válida. Além disso, verificou-se se os valores informados estavam em um formato compatível com o tipo de dado esperado. Em seguida, no conjunto de dados considerados válidos, foi realizada a deduplicação. O processo de validação e deduplicação foi implementado em

conformidade com os requisitos estabelecidos no protocolo CTDP, conforme proposto nesta tese.

Adicionalmente, foi realizada uma etapa de padronização dos dados, durante a qual os nomes próprios foram convertidos para caixa alta e os espaços em excesso foram removidos. Dessa forma, somente os dados válidos, deduplicados e padronizados foram armazenados no banco de dados. Assim, contempla-se o Objetivo C (construir uma base de dados local).

Desta forma, conclui-se o processo de estudo de praticabilidade e eficácia do protocolo CTDP proposto neste estudo, conduzido entre novembro de 2024 e junho de 2025. Os resultados obtidos e suas implicações serão detalhadamente discutidos no [Capítulo 7](#), onde são apresentados os resultados obtidos.

Ademais, mediante a construção da base de dados sobre patentes depositadas no Brasil, surge a possibilidade de validação e certificação das informações sobre patentes declaradas pelos pesquisadores em seus currículos na Plataforma Lattes. Na seção subsequente, é detalhado o processo empregado de certificação destes dados.

## 5.4 Certificação do conjunto de dados coletados

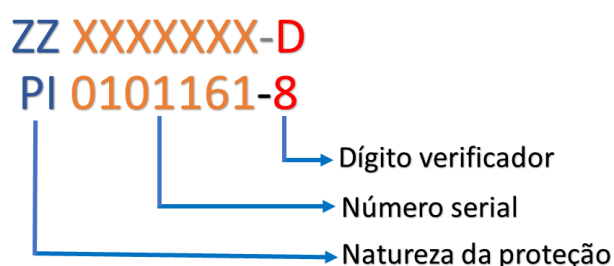
A Plataforma Lattes é amplamente reconhecida como uma fonte valiosa de informações para variados tipos de análise. Contudo, devido ao grande volume de dados e à responsabilidade do próprio indivíduo em preenchê-los, podem surgir inconsistências nos dados inseridos ([DIGIAMPIETRI et al., 2012](#)). Isso ressalta a importância de mecanismos de validação das informações contidas nos currículos da Plataforma Lattes, visando garantir a confiabilidade e certificação dessas informações. Um exemplo desse tipo de certificação é realizado pelo Oasisbr - Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto, uma iniciativa conjunta entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que certifica as teses e dissertações informadas nos currículos da Plataforma Lattes ([BRASIL, 2022](#)).

De forma análoga, uma das contribuições deste estudo é a certificação das patentes mencionadas nos currículos cadastrados na Plataforma Lattes, confrontando-as com informações registradas em repositórios internacionais, como o Espacenet. Esse processo de certificação envolve a identificação da patente registrada pelo pesquisador em seu currículo na Plataforma Lattes, verificando sua correspondência com o conjunto de patentes recuperado no repositório Espacenet. A identificação baseia-se em critérios como o número de depósito e o título da patente, garantindo, assim, maior confiabilidade nas informações apresentadas.

Destaca-se que o número de depósito de uma patente é fundamental para sua identificação tanto no INPI quanto no Espacenet, pois cada patente possui um número de depósito exclusivo. Entretanto, é importante observar que a composição desse número no INPI segue dois formatos distintos: um formato antigo, utilizado para patentes registradas em datas anteriores, e um formato mais recente, adotado desde 2 de janeiro de 2012 para novos pedidos de patente (abrangendo invenções, modelos de utilidade, desenhos industriais e indicações geográficas). Esse novo formato foi implementado para alinhar-se aos padrões internacionais de identificação de patentes, sendo o mesmo adotado pelo Espacenet (UECE, 2021).

O formato atribuído às patentes depositadas no INPI até 31/12/2011 segue a estrutura **ZZ XXXXXXX-D**, em que **ZZ** indica a natureza da proteção, **XXXXXXX** é um número serial anual composto por sete dígitos, e **D** representa o dígito verificador. A Figura 27 ilustra graficamente a composição desse formato.

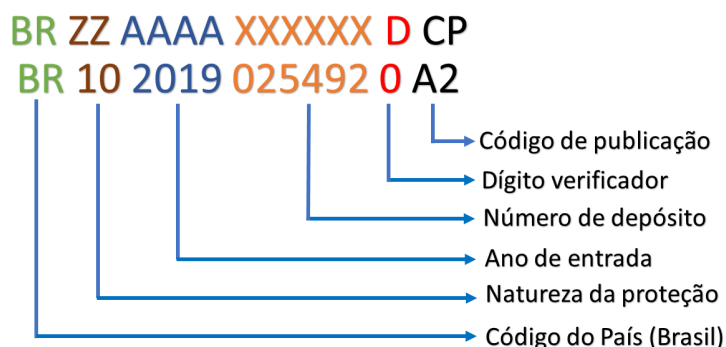
Figura 27 – Formato número de depósito de patentes adotado pelo INPI até 12/2011



Fonte: Elaboração do autor

O novo formato de numeração foi estabelecido com o objetivo de alinhar o INPI à política de integração internacional, em conformidade com os padrões recomendados pela OMPI. Esse formato apresenta a seguinte estrutura: **BR ZZ AAAA XXXXXX D CP**, onde **BR** indica o país de origem; **ZZ** refere-se à natureza da proteção; **AAAA** representa o ano de entrada no INPI; **XXXXXX** corresponde à ordem de depósito dos pedidos, composta por seis dígitos; **D** é o dígito verificador; e, por fim, **CP** indica o código de publicação, representando o status legal do pedido junto ao INPI (Apêndice B). A Figura 28 apresenta graficamente a composição do formato atual de numeração de pedidos de depósito de patentes no INPI.

Figura 28 – Formato atual de número de depósito de patentes adotado pelo INPI



Fonte: Elaboração do autor

Diante do exposto, todas as patentes depositadas no INPI até dezembro de 2011 e disponíveis na Espacenet tiveram sua numeração ajustada para atender ao padrão internacional. Assim, uma mesma patente pode possuir um número de depósito no INPI e um número diferente na Espacenet. Um exemplo dessa variação numérica pode ser observado na patente "PI 0000061-2 B1", conforme ilustrado na [Figura 29](#)..

Figura 29 – Comparativo entre o número de depósito informado no INPI e na Espacenet

BRASIL Acesso à informação Participe Serviços Legislação Canais

Instituto Nacional da Propriedade Industrial  
Ministério da Economia

Consulta à Base de Dados do INPI

» Consultar por: Base Patentes | Finalizar Sessão

Patente

(11) Nº do Pedido: PI 0000061-2 B1

(22) Data do Depósito: 14/01/2000

(43) Data da Publicação: 02/10/2001

(47) Data da Concessão: 30/06/2009

(51) Classificação IPC: E06B 1/00

(54) Título: CONJUNTO DE MONTAGEM

BRPI0000061C1

INPI: PI 0000061-2 B1

ESPAENET: BR PI 0000061 C1

BRPI0000061C1

contramarco para montagem de marcos de esquadrias

Available

1. >

bibliographic data Description Claims Drawings Original document Citations Legal events Patent family

Applicants ALCOA ALUMINIO SA [BR] +

Inventors CARDOSO ANTONIO BENEDITO +

Classifications

IPC E06B1/00;

Fonte: Elaboração do autor

Posto isto, é prudente considerar que os proponentes podem informar em seus currículos as patentes utilizando tanto o formato antigo, o qual não será localizado diretamente na Espacenet, quanto o formato novo, que possibilita sua localização no repositório. Assim, tornou-se necessário definir um conjunto de regras para tratar os números de depósito das

patentes informadas nos currículos da Plataforma Lattes, permitindo, conseqüentemente, a identificação das patentes correspondentes na Espacenet. O objetivo central é assegurar a identificação das patentes mais antigas no padrão internacional adotado pela Espacenet. As regras estabelecidas são as seguintes:

1. Regra **R001**: Remover todos os caracteres referentes a formação como: espaços, pontos, vírgulas, parênteses, traços, barras, aspas. E a remoção de termos como: “SIGILO”, “PROTOCOLO”, “Nº”, “INPI”, “DEPÓSITO”, “EM ANDAMENTO”. Exemplo de aplicação:

**(BR) 10.2019 0254920-A2 → BR1020190254920A2**

2. Regra **R002**: Aplicar a regra **R001**, caso o número inicie com “BR”, deverá ser desconsiderados os 3 últimos caracteres do número de depósito da patente. Exemplo de aplicação:

**BR 10 2019 025492 0 A2 → BR102019025492**

3. Regra **R003**: Aplicar a regra **R001**, Caso o número inicie com alguns dos seguintes prefixos: “C1”, “DI”, “MI”, “MU” ou “PI”, deverão ser desconsiderados os 3 últimos caracteres do número de depósito da patente. Exemplo de aplicação:

**PI 0000061-2 B1 → PI0000061.**

4. Regra **R004**: Aplicar a regra **R001**, Caso o número inicie com alguns dos seguintes prefixos: “C1”, “DI”, “MI”, “MU” ou “PI”. Aplicar regra **R003** e incluir o prefixo “BR”. Exemplo de aplicação:

**PI 0000061-2 B1 → BRPI0000061.**

5. Regra **R005**: Aplicar a regra **R001**, Caso o número inicie com alguns dos seguintes prefixos: “C1”, “DI”, “MI”, “MU” ou “PI”. Aplicar regra **R003** e substitui os dois primeiros caracteres do número de depósito pelo o prefixo “BR”. Exemplo de aplicação:

**PI0300001-0 A2 → BR0300001.**

6. Regra **R006**: Aplicar a regra **R001**, Caso o número não inicie com alguns dos seguintes prefixos: “BR”, “C1”, “DI”, “MI”, “MU” ou “PI”. deverão ser desconsiderados os 3 últimos caracteres do número de depósito da patente. Exemplo de aplicação:

**0.301.862-8 → 03018.**

7. Regra **R007**: Aplicar a regra **R001**, Caso o número não inicie com alguns dos seguintes prefixos: “BR”, “C1”, “DI”, “MI”, “MU” ou “PI”.Aplicar regra **R006** e incluir o prefixo “BR”. Exemplo de aplicação:.. Exemplo de aplicação:

**0.301.862-8 → BR03018.**

Após a definição das regras, foi estabelecido o fluxo de execução delas. Na [Figura 30](#) é apresentada a ordem de execução das regras definidas. No entanto, é preciso esclarecer alguns termos adotados:

- O primeiro é “busca exata”, que consiste em realizar uma busca por patente na base de dados local, através de seu número de depósito. Somente a patente que possuir exatamente o mesmo número informado como parâmetro de busca é considerada como resultado para a busca, ou seja, sucesso.
- O segundo é “busca aproximada”, que consiste em realizar uma busca por patente na base de dados local através de seu número de depósito, desta vez, desconsiderando quaisquer outros caracteres que possam existir após o parâmetro usado para consulta. Por exemplo, ao realizar uma busca usando “BRPI00” como parâmetro de busca, quaisquer registros que possuírem “BRPI00” em seu número de depósito, será considerado como resultado para a busca realizada, mesmo que existam outros caracteres após o parâmetro. Logo, patentes com números de depósito “BRPI00000061”, ou “BRPI00000023A” serão consideradas como resultado para busca aproximada por “BRPI00”.
- O terceiro termo é “sucesso”. É considerado sucesso quando, ao realizar buscas exata ou aproximada, é localizada uma ou mais patentes na base de dados local proposta neste estudo.



Diante do exposto, após localizar uma ou mais patentes, realiza-se a validação dos respectivos títulos. Esse processo consiste em calcular a distância de Levenshtein<sup>13</sup> entre o título da patente localizada e o título informado no currículo do proponente. Para aumentar a precisão, são removidas palavras muito comuns do idioma português, conhecidas no campo de Processamento de Linguagem Natural (PLN) como *stopwords*, além das acentuações. A patente é considerada válida e certificada se a similaridade calculada for igual ou superior a 85%.

O percentual de 85% foi estabelecido considerando que o preenchimento do currículo é realizado manualmente pelo autor, o que pode resultar em variações de grafia do título para a mesma patente. Um limiar inferior a 85% poderia incluir patentes não certificadas, enquanto um limite muito elevado excluiria patentes potencialmente certificáveis. A título de ilustração, a [Tabela 7](#) apresenta os resultados de algumas comparações entre títulos de patentes.

Tabela 7 – Similaridade de títulos usando Levenshtein

| Título na Plataforma Lattes                                                                                                                                                                                                                      | Título na Espacenet                                                                                                                                                                                                                                                    | Similaridade |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|
| Processo de separação do xilitol por adsorção em colunas de leito fixo de zeólitas                                                                                                                                                               | Processo de separação do xilitol por adsorção em colunas de leito fixo de zeólitas                                                                                                                                                                                     | 100%         |
| PRODUTO FERTILIZANTE SÓLIDO À BASE DE BORÓ, CÁLCIO E MAGNÉSIO, PROCESSO DE PRODUÇÃO E MODO DE APLICAÇÃO                                                                                                                                          | produto fertilizante sólido À base de boro, cálcio e magnésio, processo de produção e modo de aplicação                                                                                                                                                                | 97%          |
| BR 102015004330-9 A2 - Processo de obtenção de um extrato padronizado de mesocarpo dos frutos de <i>Atallea speciosa</i> Mart. ex Spreng (babaçu) e de formulação cosmética antioxidante para prevenção e/ou atenuação do envelhecimento cutâneo | processo de obtenção de um extrato padronizado de mesocarpo dos frutos de <i>atallea speciosa</i> mart. ex spreng (babaçu) e de formulação cosmética contendo o extrato padronizado com atividade antioxidante para prevenção e/ou atenuação do envelhecimento cutâneo | 85%          |
| Dispositivo Bipartido (Clamp) Para Aplicação de Resina em Reparo de um Duto Flexível (Riser) e/ou Monitoramento de Duto Rígido (Riser)                                                                                                           | dispositivo bipartido para aplicação de resina em reparo de um duto flexível e/ou monitoramento de duto rígido                                                                                                                                                         | 88%          |
| Uso de novos compostos capazes de absorver radiação ultravioleta, composição contendo os mesmos e processos para sua preparação.                                                                                                                 | compostos capazes de absorver radiação ultravioleta, composições contendo os mesmos e processos para sua preparação                                                                                                                                                    | 90%          |

Com base nas estratégias apresentadas, foi possível implementar um processo de certificação que visa garantir a validade dos dados de patentes inseridos na Plataforma Lattes, contribuindo com a construção de uma base de dados mais consistente, fortalecendo a integridade e a credibilidade dos currículos cadastrados.

<sup>13</sup> A distância de Levenshtein é uma métrica de texto usada para medir a diferença entre duas palavras ou frases.

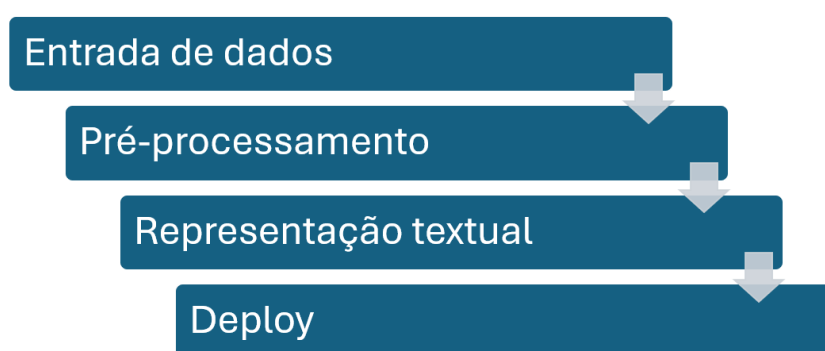
Na próxima seção são apresentadas as técnicas de inteligência artificial adotadas na ferramenta Pytente.

## 5.5 Pipeline de aprendizado de máquina e processamento de linguagem natural

Esta seção tem como objetivo detalhar as técnicas de inteligência artificial, com ênfase em *Machine Learning* e Processamento de Linguagem Natural (PLN), incorporadas na ferramenta Pytente, desenvolvida no âmbito desta tese. Assim, amparando o Objetivo D (propor uma estratégia baseada em técnicas de *Large Language Models* (LLM) para viabilizar buscas semânticas, com o intuito de aprimorar a compreensão da linguagem técnica presente nos documentos de patentes).

O *pipeline* implementado adotado representa um *workflow* estruturado, ou seja, uma sequência de etapas que abrange desde a entrada dos dados até a obtenção dos resultados finais (Figura 31).

Figura 31 – Pipeline de *Machine Learning* e Processamento de Linguagem Natural



Fonte: Elaboração do autor

Conforme apresentado na figura, o processo tem início na camada de entrada de dados, na qual a fonte é composta pelo conjunto de patentes únicas armazenadas na base de dados local estabelecida nesta tese. Na camada de pré-processamento, são realizadas etapas fundamentais para a preparação dos dados, como a **Tokenização**, que consiste na divisão do texto em unidades menores chamadas tokens. Esses tokens incluem todas as palavras, pontuações e símbolos presentes no título ou resumo das patentes, conforme a necessidade da análise em questão. Após a tokenização, ocorre a etapa de **Remoção de Pontuação**, na qual todos os tokens compostos exclusivamente por caracteres de pontuação são removidos, uma vez que não contribuem para a interpretação semântica do texto.

Por fim, ainda na camada de pré-processamento, ocorre a **Remoção de Stop Words**, processo no qual são excluídas palavras de alta frequência e baixo valor informativo, como [a], [o] e [de]. A eliminação dessas palavras é importante para evitar impactos negativos no desempenho do modelo, visto que, assim como a pontuação, não agregam significado relevante ao contexto analisado.

Na camada de representação textual, são aplicadas técnicas que permitem a interpretação dos tokens pelo modelo. Uma das abordagens adotadas é a métrica **TF-IDF** (*Term Frequency-Inverse Document Frequency*), que atribui pesos aos termos de um documento com base em sua frequência local e global. Dessa forma, termos recorrentes em um documento específico, mas raros no conjunto total, recebem pesos mais elevados, contribuindo para uma representação mais eficiente. Outra técnica utilizada é o **Word2Vec**, um algoritmo que converte palavras em vetores numéricos em um espaço de baixa dimensão, estabelecendo relações semânticas entre os termos.

A base computacional foi estruturada a partir do modelo **BERT** (*Bidirectional Encoder Representations from Transformers*), desenvolvido pela Google, classificado como uma LLM (*Large Language Model*), ou seja, um modelo de aprendizado profundo treinado com grandes volumes de dados textuais. Além disso, foi adotada a variante S-BERT (*Sentence-BERT*), uma extensão do BERT, otimizada para a geração de embeddings semânticos de sentenças. Essa implementação é viabilizada por meio da biblioteca Python SentenceTransformers<sup>14</sup>.

A biblioteca SentenceTransformers permite o cálculo de representações vetoriais (*embeddings*) de frases e textos, sendo especialmente eficaz na identificação de sinônimos e na aproximação semântica entre descrições textuais, o que a torna adequada para tarefas de busca em bases documentais complexas, como a de patentes. Além disso, o modelo oferece suporte a mais de 100 idiomas, incluindo o português brasileiro, o que garante maior aderência às particularidades linguísticas da base nacional de patentes. O repositório oficial da ferramenta S-BERT disponibiliza uma variedade de modelos pré-treinados, otimizados para diferentes tarefas de PLN, e pode ser acessado em <<https://www.sbert.net/index.html>>.

Feito isso, a base necessária para viabilizar estratégias de Inteligência Artificial no conjunto de dados de patentes está estabelecida para criar clusterização de patentes a partir de seus resumos. Foi utilizada a biblioteca BERTopic<sup>15</sup>, amplamente conhecida por sua relevância na modelagem de tópicos em grandes volumes de texto. Essa biblioteca combina os princípios do modelo BERT com a eficácia do c-TF-IDF, permitindo a criação de agrupamentos de diversos documentos de patentes.

Uma das principais vantagens da adoção do BERTopic é a disponibilidade de modelos previamente treinados em português, o que reduz a necessidade de treinamentos.

<sup>14</sup> <<https://pypi.org/project/sentence-transformers>>

<sup>15</sup> <<https://maartengr.github.io/BERTopic/index.html>>

Além disso, a biblioteca é capaz de lidar com grandes volumes de dados, abrangendo tanto textos curtos quanto longos, e oferece ferramentas avançadas de visualização, que facilitam a exploração e interpretação dos resultados obtidos.

Já para a implementação de busca semântica no conjunto local de patentes, foi adotado o modelo pré-treinado “distiluse-base-multilingual-cased-v1”, pertencente à família sentence-transformers. Este modelo mapeia sentenças e parágrafos para um espaço vetorial denso de 512 dimensões, sendo particularmente adequado para tarefas como agrupamento de documentos e recuperação semântica de informações (REIMERS; GUREVYCH, 2019).

A implementação da busca semântica de patentes requer uma etapa inicial de preparação do banco de dados, na etapa de pré-processamento de dados, que visa estruturar as informações em um formato compatível para a análise semântica. Neste estudo, foram considerados apenas os campos título e resumo das patentes, uma vez que são os elementos que concentram as descrições mais relevantes do conteúdo técnico.

O pré-processamento, adotado na busca semântica, é dividido em quatro etapas. A primeira etapa (i) consiste na seleção estruturada das patentes da base de dados local, extraíndo os seguintes atributos: número da patente, título e resumo. Em seguida, realiza-se a etapa (ii) de limpeza de dados, na qual são excluídos todos os registros que não contenham os campos obrigatórios (número, título ou resumo), bem como aqueles que apresentem valores inconsistentes ou inválidos para os campos classificados como obrigatórios. A etapa subsequente (iii) corresponde à transformação dos dados, que envolve a conversão de todos os caracteres para minúsculas e a remoção de palavras irrelevantes (stopwords), a fim de reduzir ruídos semânticos. Finalmente, na etapa (iv) é feita a redução dos dados, em que os campos “título” e “resumo” são concatenados em uma única descrição, resultando em uma base tabular contendo apenas dois atributos: número da patente e descrição consolidada.

Concluída a preparação do conjunto de dados, é iniciada a fase de processamento semântico, na qual o conjunto de dados tratado é convertido em vetores numéricos (embeddings), por meio do modelo “distiluse-base-multilingual-cased-v1”. Essa representação vetorial captura o significado das sentenças e suas relações semânticas, viabilizando sua posterior análise por algoritmos de aprendizado de máquina. É importante destacar que, dada a natureza computacionalmente complexa do processo de geração de embeddings, especialmente em bases extensas, foi adotada uma estratégia de indexação por ano de depósito. Com isso, os dados são segmentados anualmente, reduzindo o volume de documentos por lote e otimizando o desempenho computacional.

Concluída as fases de pré-processamento e processamento, o conjunto de dados fica otimizado para execução de processamento semânticos como busca e agrupamentos. Para a execução da busca semântica, o texto usado para consulta, por exemplo, “sistema de descongelamento para aparelho de refrigeração”, também passa por um pipeline de

pré-processamento, incluindo a conversão para minúsculas, a remoção de stopwords e a transformação em embedding. Em seguida, o vetor resultante é comparado com os vetores das patentes indexadas por ano, utilizando medidas de similaridade semântica. O modelo SentenceTransformer retorna, para cada patente, um escore de relevância em relação ao termo consultado, variando entre 0.0 (baixa similaridade) e 1.0 (alta similaridade).

Este capítulo foi dedicado à apresentação do desenvolvimento e validação das estratégias de coleta e tratamento de dados de patentes, conforme o protocolo proposto nesta tese, além de novas abordagens para a certificação de dados e análise de patentes. Durante o processo, foi possível construir uma base de dados local composta por informações de patentes, programas de computador com registro e currículos da Plataforma Lattes, resultando na integração de três importantes fontes de dados: Espacenet, INPI e Plataforma Lattes. Essa junção caracteriza uma alternativa viável para as dificuldades enfrentadas na coleta de grandes volumes de dados de patentes.

Ao final deste capítulo, fica claro que as estratégias de coleta e tratamento de dados propostas utilizam ferramentas amplamente consolidadas no meio acadêmico e profissional.

Nos próximos capítulos, são apresentados os resultados obtidos ao longo do desenvolvimento desta tese, bem como a caracterização da base de dados construída por meio das estratégias descritas neste estudo.

## 6 Pytente

Este capítulo tem como objetivo principal apresentar um dos resultados relevantes desta tese, decorrente da implementação do Protocolo de Coleta e Tratamento de Dados Bibliográficos de Patentes (Protocolo CTDp), proposto neste estudo. Destaca-se, nesse contexto, o desenvolvimento do conjunto ferramental denominado “Pytente”, um *framework* projetado para automatizar as diretrizes estabelecidas pelo Protocolo CTDp.

A implementação desta ferramenta revela a aplicabilidade do Protocolo CTDp, bem como a caracterização de suas limitações e os desafios enfrentados ao longo de sua implementação. Na sequência, são detalhados os resultados do estudo de praticabilidade e eficácia do protocolo, conforme exposto na [Seção 6.1](#), além de serem descritas, na [Seção 6.2](#), as informações técnicas relativas ao Pytente.

### 6.1 Estudo de praticabilidade e eficácia do protocolo CTDp

O estudo de praticabilidade e eficácia, discutido anteriormente na [Seção 5.3](#), tem como objetivo avaliar a eficiência do protocolo CTDp na coleta e tratamento de grandes volumes de dados de patentes. O que resultou no desenvolvimento da ferramenta Pytente, criada para automatizar todo o processo.

No estudo de praticabilidade e eficácia, o protocolo CTDp foi aplicado com o objetivo de coletar patentes depositadas no Brasil, no período de 1900 a 2024, visando recuperar um conjunto amplo e representativo de dados. Os resultados demonstraram a eficácia do protocolo na recuperação de informações, a partir de repositórios de acesso aberto, respeitando integralmente as diretrizes de coleta estabelecidas pelo repositório de patentes utilizado, neste caso, o Espacenet.

Contudo, a aplicação do protocolo em outros repositórios de patentes exige a implementação de tratamentos específicos. Isso ocorre devido às variações na estrutura dos dados entre diferentes plataformas, o que inviabiliza a adoção de um procedimento genérico. Assim, torna-se necessário o desenvolvimento de regras de coleta e tratamento customizadas para cada repositório, garantindo a compatibilidade e a integridade dos dados coletados.

Na primeira versão do protocolo CTDp, o repositório Espacenet foi definido como a principal fonte de coleta, com base na aplicação da “Matriz de triagem de Repositório de Patentes” desenvolvida neste estudo. 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.

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, em um ambiente computacional com 16 GB de memória RAM e processador Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz. Essa estratégia permitiu a coleta integral do conjunto de dados em apenas 7 meses, demonstrando a eficácia da 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.

Diante dos resultados obtidos, a aplicação do protocolo CTDP com a ferramenta Pytente se mostra uma alternativa viável para a coleta de grandes conjuntos de dados de patentes.

No contexto da Ciência Aberta, o Pytente apresenta-se como uma solução tecnológica que amplia significativamente o acesso, a transparência e a reprodutibilidade das pesquisas relacionadas à propriedade industrial e tecnológica. Sua capacidade de automatizar a coleta de grandes volumes de dados bibliográficos de patentes, armazená-los de forma estruturada e realizar procedimentos de validação e deduplicação facilita a formação de bases de dados locais consistentes e atualizados, alinhados aos princípios de acesso aberto aos dados científicos.

Sua natureza flexível possibilita aplicação em diversos escopos de pesquisa envolvendo documentos de patentes, atendendo às demandas de pesquisadores, gestores de inovação, núcleos de inovação tecnológica (NITs) e demais atores do ecossistema científico e tecnológico. Essa característica permite a seleção e análise de dados específicos de acordo com as necessidades de cada usuário, sem a exigência de infraestrutura computacional avançada ou conhecimento aprofundado em mineração de dados. Assim, o Pytente reforça os princípios FAIR (*Findable, Accessible, Interoperable e Reusable*), garantindo que os dados científicos sejam encontráveis, acessíveis, interoperáveis e reutilizáveis.

De maneira geral, o Pytente atua como um facilitador que transforma informações já disponíveis em repositórios abertos em recursos ainda mais acessíveis e utilizáveis, que é

a essência da Ciência Aberta. Nesse sentido, destaca-se sua contribuição em diferentes dimensões:

**Aceleração do Acesso a Dados Abertos:** ao automatizar a coleta de dados bibliográficos de patentes em repositórios de acesso aberto, o Pytente agiliza o processo de recuperação da informação, tornando os dados de patentes rapidamente disponíveis para análise e evitando esforços manuais extensos. Essa automação reduz o tempo de busca e amplia a disponibilidade de dados relevantes à pesquisa e à inovação.

**Promoção da Reutilização e Validação de Dados:** ao garantir o armazenamento estruturado, a validação e a eliminação de registros duplicados, o Pytente assegura a qualidade e a confiabilidade dos dados, requisitos essenciais para a Ciência Aberta. Dados limpos e organizados potencializam sua reutilização em novas análises, validação de resultados e desenvolvimento de pesquisas subsequentes, ampliando o ciclo de geração e uso de conhecimento científico.

**Facilitação da Descoberta e Inovação:** suas funcionalidades de busca semântica, que permitem a recuperação de informações, por meio de linguagem natural sem necessidade de expressões lógicas complexas, além da extração personalizada de subconjuntos de dados, reduzem barreiras técnicas para a exploração de patentes. Essa abordagem torna as informações tecnológicas mais acessíveis a pesquisadores de diferentes áreas do conhecimento, incluindo Ciências Humanas, Ciências Sociais Aplicadas e Saúde, fomentando a interdisciplinaridade e a inclusão digital.

Ao possibilitar a criação de bases de dados locais derivadas de repositórios de acesso aberto, o Pytente favorece a integração de seus dados a sistemas de informação científica, repositórios institucionais, plataformas de gestão da pesquisa e ferramentas de análise de dados. Essa integração potencializa análises bibliométricas, patentométricas e estudos prospectivos, fortalecendo a construção de ecossistemas digitais de pesquisa abertos. Consequentemente, contribui para o desenvolvimento de inovações alinhadas a demandas sociais e econômicas, e para a ampla disseminação do conhecimento técnico-científico em benefício da sociedade.

Conforme apresentado, para além de uma inovação tecnológica no processo de coleta e análise de dados bibliográficos de patentes, a criação do Pytente representa um compromisso intrínseco com os princípios da Ciência Aberta e da democratização do acesso à informação tecnológica. Assim, para que sua adoção e aplicabilidade sejam efetivamente ampliadas e perdurem no médio e longo prazo, é importante estabelecer um plano de sustentabilidade, que contemple aspectos técnicos, institucionais, comunitários e financeiros.

O licenciamento do Pytente, sob uma licença de software livre, como a licença MIT<sup>1</sup>,

<sup>1</sup> A Licença MIT, desenvolvida pelo Massachusetts Institute of Technology (MIT), é uma licença de software

é uma escolha estratégica que possibilita sua livre utilização, modificação e redistribuição por pesquisadores, Núcleos de Inovação Tecnológica (NITs), universidades e centros de pesquisa. Essa abordagem favorece a construção de uma comunidade ativa de usuários e desenvolvedores e se alinha diretamente aos preceitos da Ciência Aberta, garantindo a transparência e a reprodutibilidade dos processos computacionais inerentes à ferramenta.

A sustentabilidade técnica do Pytente é viabilizada, principalmente pela publicação do seu código-fonte no repositório colaborativo GitHub (<[https://github.com/raulivan/pytent\\_library](https://github.com/raulivan/pytent_library)>). Essa prática assegura o controle de versionamento e facilita contribuições externas, correções de erros e a implementação de melhorias contínuas, por parte da comunidade de pesquisadores e desenvolvedores. A plataforma GitHub, por sua natureza, atua como um hub centralizado que promove a transparência do desenvolvimento garantindo a evolução e a adaptabilidade da ferramenta às novas demandas e tecnologias.

Para consolidar o Pytente como uma ferramenta de uso estratégico, é essencial a articulação com grupos de pesquisa, redes de colaboração científica e instituições públicas. A incorporação da ferramenta em projetos de pesquisa, programas de pós-graduação e núcleos de apoio à inovação tecnológica tem o potencial de ampliar exponencialmente sua base de usuários e gerar um fluxo contínuo de solicitações de melhorias e novas funcionalidades. Adicionalmente, a elaboração e oferta regular de oficinas, tutoriais, cursos livres e minicursos em eventos científicos são ações cruciais para a divulgação e adoção do ferramental proposto. Estas iniciativas contribuem para reduzir as barreiras de entrada, estimulam o uso efetivo da ferramenta e, principalmente, fortalecem a autonomia de pesquisadores na coleta e análise de dados de patentes.

Na próxima seção, são apresentadas informações técnicas de instalação, configuração e uso do Pytente.

## 6.2 Detalhamento técnico e procedimentos de uso do Pytente

O conjunto ferramental “Pytente”, é um *framework* projetado para automatizar as diretrizes estabelecidas pelo Protocolo CTDP também proposto nesta Tese. O Pytente pode ser facilmente acessado pelo endereço eletrônico <<https://pytentui.web.app/>> (Figura 32).

---

de caráter permissivo, amplamente utilizada tanto em projetos de código aberto quanto em aplicações proprietárias.

Figura 32 – Página inicial do Pytente



Fonte: Elaboração do autor

O Pytente não opera integralmente em um servidor remoto, é necessário realizar também a instalação em um computador local para seu pleno funcionamento. Sua interface web permite a construção de estratégias para coleta, tratamento e extração de subconjuntos de dados. No entanto, a execução dessas estratégias ocorre localmente, conferindo ao pesquisador maior autonomia sobre todo o processo. Para melhor acessibilidade ao uso da ferramenta, a interface web pode ser acessada nos idiomas Português do Brasil e Inglês.

O Pytente foi desenvolvido para ser compatível com os sistemas operacionais Windows, Linux e macOS, exigindo apenas que o usuário tenha previamente instalado o Python<sup>2</sup>, 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. Todo o processo de instalação, configuração e principais funcionalidades do Pytente é descrito no [Apêndice C](#).

No próximo capítulo são apresentadas análises feitas mediante o uso da ferramenta Pytente, desde o processo de coleta até a extração de subconjuntos de dados.

<sup>2</sup> <<https://www.python.org/downloads/>>

# 7 Análise e Discussão dos Resultados

Este capítulo tem como objetivo apresentar e discutir os resultados obtidos a partir da coleta de dados bibliográficos de patentes, realizada com base nas técnicas propostas pelo Protocolo CTD<sub>P</sub> desenvolvido neste estudo.

## 7.1 Apresentação e descrição dos dados

Com o objetivo de responder à questão de pesquisa proposta nesta tese, foram coletados dados bibliográficos sobre patentes utilizando as estratégias delineadas ao longo deste estudo. Esta seção destina-se a apresentar e descrever o conjunto de dados estabelecido, fornecendo um panorama geral de sua estrutura e conteúdo, bem como destacando sua relevância para os objetivos desta pesquisa.

Nesse contexto, e em termos quantitativos, a [Tabela 8](#) ilustra o volume de patentes recuperadas no repositório Espacenet, considerado o conjunto de dados principal da base desenvolvida neste estudo.

Tabela 8 – Quantitativos de patentes coletadas no repositório Espacenet

| Tipo de registro  | Fonte     | Sigla | Quantidade |
|-------------------|-----------|-------|------------|
| Patentes - Global | Espacenet | PG    | 946.367    |
| Patentes - Único  | Espacenet | PU    | 831.706    |

Conforme apresentado na [Tabela 8](#), foi possível construir uma base de dados composta por 946.367 documentos de patentes provenientes do repositório Espacenet, referenciados pela sigla PG. Considerando apenas a última publicação de cada patente, obteve-se um total de 831.706 patentes (PU). As siglas foram adotadas com o intuito de facilitar futuras referências aos conjuntos apresentados. O termo "global" refere-se ao conjunto completo de registros, enquanto o termo "único" se aplica exclusivamente aos registros distintos.

Em dezembro de 2024, o INPI disponibilizava 963.479 patentes para consulta em seu portal de pesquisa. Esse número permite concluir que o conjunto de patentes únicas (PU), composto por 831.706 registros, representa aproximadamente 86,33% do total de patentes disponíveis no INPI em dezembro de 2024.

Entretanto, é relevante observar que o INPI não disponibiliza imediatamente as patentes no Espacenet, após sua inclusão no portal de pesquisa. Há um intervalo temporal

entre a publicação no INPI e sua inserção na base de dados do Espacenet, o que sugere que algumas das patentes disponíveis no INPI poderiam não estar acessíveis no Espacenet durante o período de coleta dos dados.

A [Tabela 10](#), apresentada a seguir, detalha o quantitativo de currículos extraídos da Plataforma Lattes que foram incorporados à base de dados local.

Tabela 10 – Quantitativos de registros da Plataforma Lattes

| Tipo de registro                                                      | Fonte             | Sigla | Quantidade |
|-----------------------------------------------------------------------|-------------------|-------|------------|
| Total de currículos com patentes e programas de computador registrado | Plataforma Lattes | CPL   | 35.072     |
| Currículos com patentes                                               | Plataforma Lattes | —     | 28.581     |
| Currículos com programas de computador registrado                     | Plataforma Lattes | —     | 8.479      |
| Patentes - Global                                                     | Plataforma Lattes | PGPL  | 76.619     |
| Patentes - Único                                                      | Plataforma Lattes | PUPL  | 57.281     |
| Programas de computador com registro - Global                         | Plataforma Lattes | PCGPL | 19.264     |
| Programas de computador com registro - Único                          | Plataforma Lattes | PCUPL | 14.708     |
| Patentes do Lattes certificadas – Global                              | Plataforma Lattes | —     | 44.856     |
| Patentes do Lattes certificadas – Único                               | Plataforma Lattes | —     | 18.678     |

Conforme apresentado, a base de dados construída, no que se refere aos currículos da Plataforma Lattes (CPL), reúne um total de 35.072 currículos. Deste montante, 28.581 incluem registros de patentes, enquanto 8.479 apresentam ocorrências de programas de computador registrados.

Em uma análise mais detalhada, os currículos da Plataforma Lattes contêm, ao todo, 76.619 registros de patentes (PGPL). Considerando apenas os registros únicos, definidos pelo critério de distinção que utiliza o título e o ano da patente, esse total é reduzido para 57.281 patentes (PUPL). No que diz respeito aos programas de computador registrados, foram identificados 19.264 registros informados nos currículos. Destes, 14.708 são registros únicos, aplicando-se os mesmos critérios de distinção baseados no título e no ano do programa (PCUPL).

Embora os currículos da Plataforma Lattes apresentem uma representatividade limitada, sua adoção como fonte complementar de dados mostrou-se valiosa, especialmente para a validação das informações sobre patentes mencionadas nos currículos. Por meio da identificação dessas patentes no conjunto de dados coletados no Espacenet, foi possível validar 44.856 registros do conjunto PGPL no conjunto PG.

Esses registros são classificados como patentes certificadas, pois as informações fornecidas nos currículos da Plataforma Lattes foram correspondidas com sucesso às patentes disponíveis no Espacenet, utilizando as estratégias metodológicas desenvolvidas neste estudo. É importante destacar que, dentre as patentes certificadas, 18.678 correspondem a registros únicos.

Os dados referentes aos depositantes e inventores, obtidos junto ao INPI, agregaram à base de dados informações sobre nacionalidade, estado e cidade de 154.291 depositantes e 605.467 inventores.

## 7.2 Análise exploratória dos dados

A análise exploratória de dados constitui uma etapa essencial no processo de pesquisa, pois permite compreender e contextualizar o conjunto de dados estabelecido, além de fornecer subsídios iniciais para investigações mais detalhadas. Nesta seção, são apresentados os resultados dessa análise exploratória, com o objetivo de descrever as características principais do conjunto de dados coletados e, assim, caracterizar a produção técnica brasileira com base nos documentos de patentes.

Nesse contexto, e com o intuito de traçar um panorama da evolução tecnológica no Brasil, foi realizada uma análise dos depósitos anuais de patentes no período compreendido entre 1970<sup>1</sup> e 2024<sup>2</sup>. Essa análise baseia-se no conjunto de patentes únicas (PU) e busca evidenciar tendências e padrões temporais.

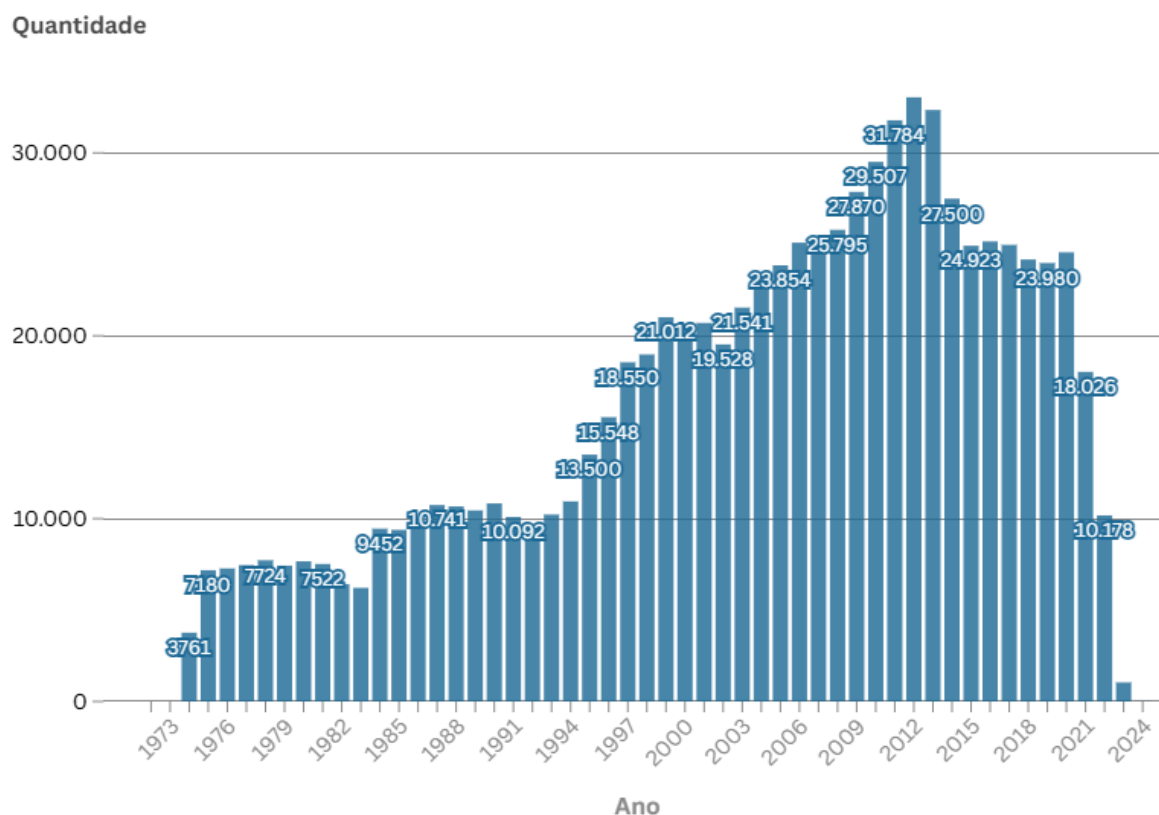
Os resultados dessa análise são ilustrados na [Figura 33](#), que apresenta a evolução temporal do total de depósitos de patentes. É importante destacar que registros de patentes que não continham a informação do ano de depósito foram excluídos da análise, a fim de assegurar a consistência dos dados apresentados.

---

<sup>1</sup> A patente mais antiga coletada na Espacenet é de 1972

<sup>2</sup> Ano limite de análise

Figura 33 – Evolução temporal do depósito de patentes por ano



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/20779400/>>

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.

As variações observadas no número de depósitos de patentes no Brasil ao longo das décadas podem ser atribuídas a uma multiplicidade de fatores de natureza econômica, política, social e tecnológica. Considerando o período de crescimento compreendido entre 1973 e 2013, é possível segmentá-lo em três fases distintas: (1) de 1973 a 1989; (2) de 1991 a 2000; e (3) de 2000 a 2013.

Na primeira fase (1973 a 1989), observa-se um crescimento moderado, porém contínuo, no volume de depósitos. Uma hipótese plausível para esse comportamento está

associada ao processo de industrialização, ao crescimento econômico e à gradual abertura comercial ocorridos nas décadas de 1970 e 1980. Mesmo diante da crise econômica, vivenciada na década de 1980, marcada por episódios de hiperinflação, é possível identificar uma tendência crescente até o final desse período. Esse fenômeno pode estar relacionado ao aumento da conscientização sobre a importância da proteção da propriedade intelectual e aos primeiros esforços de modernização do Instituto Nacional da Propriedade Industrial (INPI).

Na segunda fase (1991 a 2000), destaca-se a estabilidade econômica proporcionada pelo Plano Real em 1994, que criou um ambiente mais propício para investimentos em pesquisa, desenvolvimento e inovação. Nesse contexto, a abertura econômica iniciada nos anos 1990, aliada à adesão do Brasil ao Acordo sobre os Aspectos dos Direitos de Propriedade Intelectual Relacionados ao Comércio (TRIPS), em 1996, resultou em um aumento significativo nos depósitos de patentes, sobretudo por empresas estrangeiras interessadas em proteger suas invenções no mercado brasileiro.

A terceira fase (2000 a 2013) corresponde ao período de maior expansão no número de depósitos. Esse aumento pode ser explicado pelo crescimento econômico experimentado nos anos 2000, combinado a iniciativas governamentais de incentivo à inovação, como os editais de fomento à P&D promovidos por agências de fomento, especialmente o CNPq. Soma-se a isso, o fortalecimento do marco legal da inovação, especialmente com a promulgação da Lei de Inovação (Lei nº 10.973/2004), que ampliou o apoio institucional à proteção do conhecimento tecnológico.

Entretanto, é possível observar uma tendência de queda nos depósitos de patentes no período subsequente, de 2013 a 2024. Esse declínio, no entanto, não se reflete integralmente nos dados oficiais divulgados pelo INPI, em maio de 2024, os quais apontam uma média de 27.267,50 depósitos anuais no intervalo de 2020 a 2024 ([BRASIL, 2024](#)). Uma hipótese para essa discrepância reside na possível ausência de parte desses depósitos na base de dados da Espacenet, no momento da coleta dos dados, o que pode ocorrer devido a atrasos burocráticos no processamento e publicação das patentes ou pelo fato de alguns pedidos ainda estarem sob sigilo legal.

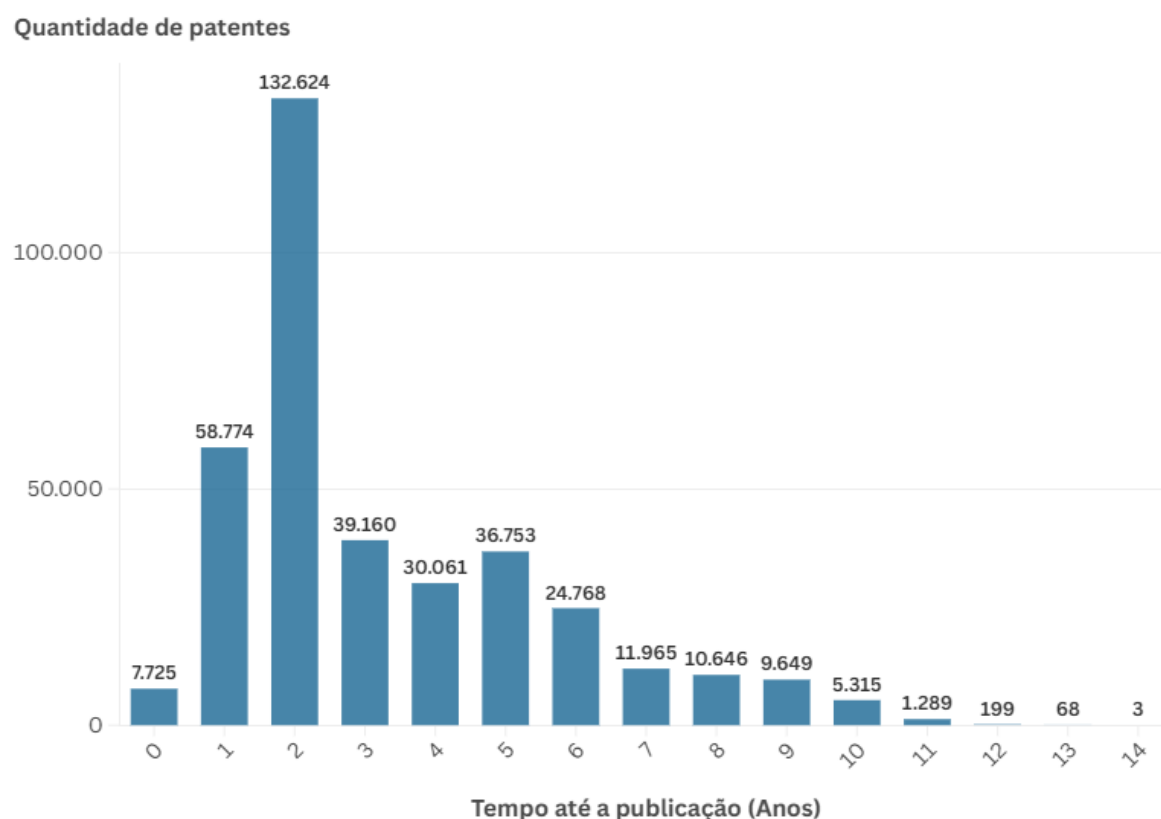
Adicionalmente, é importante considerar o contexto de instabilidade política e econômica enfrentado pelo Brasil a partir de meados de 2013. Situações de crise econômica tendem a provocar a retração dos investimentos em pesquisa e desenvolvimento, tanto por parte do setor público quanto do setor privado que, diante da necessidade de contenção de gastos, priorizam ações de curto prazo. A redução nos orçamentos destinados à ciência, tecnologia e inovação compromete a continuidade de projetos e afeta diretamente o volume de novas invenções e, conseqüentemente, o número de depósitos de patentes.

Mediante o exposto, é importante ter cuidado ao usar o Protocolo CTDP para analisar períodos recentes, pois os dados coletados podem não representar fielmente a realidade,

por isso, é importante usar fonte de dados complementares, como, neste caso, dados do INPI, para cruzar as informações.

Com o objetivo de sustentar a hipótese anteriormente apresentada, a [Figura 34](#) ilustra 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 34 – Tempo de depósito até a publicação de patentes no período de 2010 a 2024



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/20819556/>>

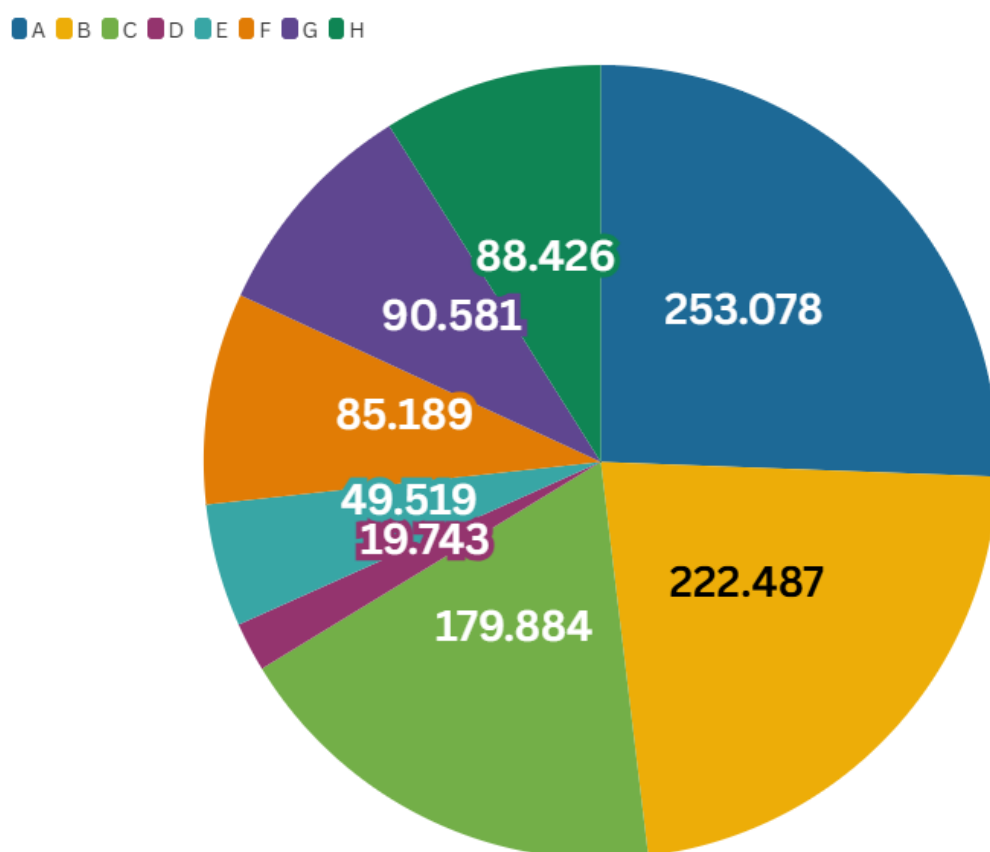
A análise revela que a maior parte dos depósitos de patentes no Brasil é publicada em um intervalo de até dois anos após o depósito. Contudo, observa-se uma variabilidade temporal, com casos extremos em que o tempo entre o depósito e a publicação efetiva atinge até 14 anos.

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 números apresentados na [Figura 33](#), subestimando o total de patentes depositadas.

A análise dos fatores que contribuíram para o aumento no número de depósitos de patentes é fundamental para compreender o panorama apresentado. Nessa perspectiva, a avaliação dos setores econômicos com maior número de depósitos permite identificar aqueles mais inovadores e com maior potencial de crescimento.

Uma abordagem relevante para essa análise baseia-se nas classificações atribuídas às patentes. Cada patente, conforme sua natureza e finalidade, é categorizada de acordo com o sistema internacional de classificação de patentes. Essas classificações fornecem informações sobre as áreas tecnológicas predominantes e permitem compreender quais setores têm sido mais prolíficos em gerar novas patentes. A [Figura 35](#) apresenta as classificações com maior número de registros, oferecendo uma visão detalhada dos domínios tecnológicos mais representativos. Essa perspectiva destaca os setores mais ativos em inovação.

Figura 35 – Total de patentes pela a primeira classificação atribuída no processo de depósito



Fonte: Elaboração do autor

A – Necessidades Humanas; B – Operações de Processamento; Transporte; C – Química e Metalurgia; D – Têxteis e Papel; E – Construções Fixas; F – Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão; G – Física; H – Eletricidade

Acesso online: <<https://public.flourish.studio/visualisation/20845696/>>

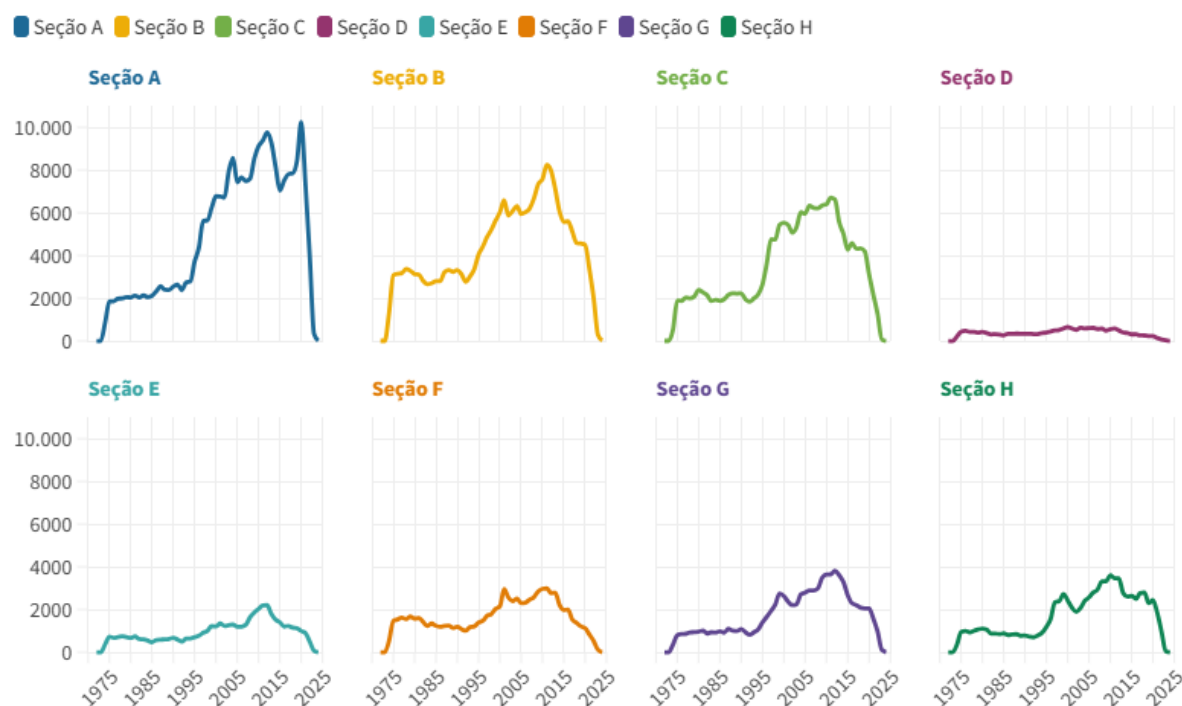
O gráfico apresentado na [Figura 35](#) foi elaborado com base nos dados do conjunto PG, levando em consideração possíveis alterações nas classificações atribuídas às patentes ao longo do tempo.

De forma geral, observa-se que a maior parte das patentes brasileiras está concentrada em três classificações: A – Necessidades Humanas (25,60%), B – Operações de Processamento e Transportes (22,50%) e C – Química, Metalurgia (18,19%), juntas representam 66,29% das patentes depositadas no Brasil. Essas categorias abrangem setores estratégicos para a economia brasileira, como agricultura, indústria, serviços, tecnologia da informação, energia e mineração. Esses setores desempenham papéis fundamentais na economia nacional, sendo motores de geração de empregos, fontes significativas de receita e catalisadores para o desenvolvimento econômico e tecnológico do país ([CNA, 2021](#)).

Nesta perspectiva, é possível considerar que a elevada concentração de patentes em setores como agricultura, indústria e energia evidencia um esforço direcionado ao fortalecimento de áreas estratégicas para o desenvolvimento socioeconômico do país.

A [Figura 36](#) apresenta, em uma análise temporal, a evolução do número de depósitos de patentes ao longo dos anos, distribuídos entre as oito seções que compõem o sistema internacional de classificação de patentes.

Figura 36 – Total de depósito de patentes por ano e por classificação



Fonte: Elaboração do autor

A – Necessidades Humanas; B – Operações de Processamento; Transporte; C – Química e Metalurgia; D – Têxteis e Papel; E – Construções Fixas; F – Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão; G – Física; H – Eletricidade

Acesso online: <<https://public.flourish.studio/visualisation/14042678/>>

Conforme apresentado na Figura 36, algumas seções apresentam um crescimento relevante no número de depósitos de patentes ao longo dos anos (Seções A, B, C, G e H), indicando um aumento significativo de interesse e inovação nessas áreas. Por outro lado, as Seções E e F apresentam um padrão mais estável, com menor variação, refletindo que a atividade de patenteamento nessas áreas se mantém constante ao longo do tempo. Já na Seção D, observa-se uma redução no número de depósitos, o que pode refletir uma possível saturação do mercado ou alterações nas prioridades de pesquisa e desenvolvimento.

A partir da década de 1970, as áreas relacionadas a operações de processamento e transporte (Seção B) se destacaram, liderando o número de depósitos de patentes até 1995. Contudo, após esse período, as patentes classificadas na “Seção A - Necessidades Humanas” apresentaram maior número de patentes depositadas. O crescimento dessa seção pode indicar o surgimento de novas tecnologias ou mudanças nos incentivos que estimulam a inovação nessa área, resultando em uma intensa atividade de pesquisa e desenvolvimento.

Adicionalmente, um padrão de crescimento mais estável ou até mesmo uma tendência de declínio, como o observado na Seção D, pode refletir setores mais maduros, nos quais a tecnologia já está consolidada e a inovação se concentra em aprimoramentos.

Conforme destacado anteriormente na [Capítulo 2](#), as classificações atribuídas às patentes possuem uma estrutura hierárquica organizada em cinco níveis: Seção, Classe, Subclasse, Grupo e Subgrupo, citados do mais abrangente ao mais restritivo. Essa estrutura permite agrupar as patentes, facilitando a análise de suas características e áreas de aplicação.

Nesse contexto, avaliar a distribuição das classificações entre as classes que compõem uma seção viabiliza uma visão detalhada dos diferentes grupos tecnológicos que compõem a classificação e permite identificar se as patentes estão bem distribuídas entre as diferentes categorias. Esta abordagem permite compreender a concentração ou dispersão das inovações dentro de uma seção específica, revelando tendências e áreas com maior ou menor atividade de patenteamento. A [Figura 38](#) apresenta a distribuição das patentes em cada seção.



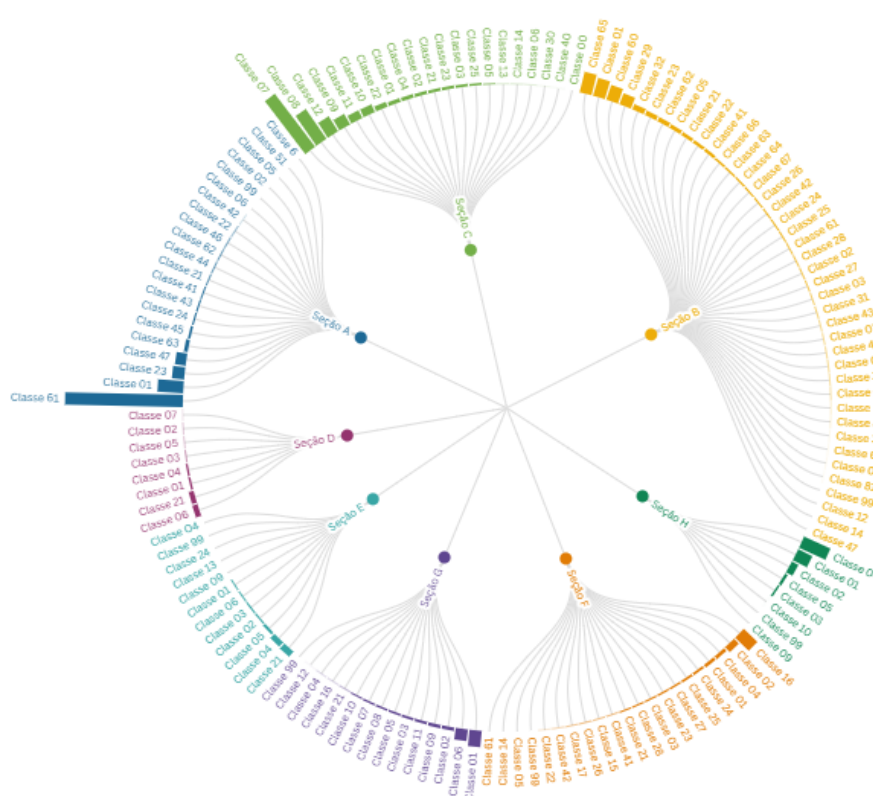
classes) e F (21 classes) apresentam a maior distribuição de classes. Esse alto número de classes pode ser interpretado como um indicativo de áreas com intensa atividade inventiva ou grande interesse tecnológico, destacando a complexidade e o desenvolvimento avançado das tecnologias associadas a essas seções. Isso reflete setores economicamente estratégicos e altamente competitivos. Como já destacado anteriormente, as seções A, B e C, em conjunto, representam mais da metade das patentes depositadas no Brasil.

Por outro lado, a Seção D e H são as que apresentam a menor distribuição de classes, com apenas 8 classes cada. Esse fato pode ser interpretado de duas perspectivas, (1) uma seção com poucas classes e baixa atividade inventiva, como a Seção D, pode sugerir que as tecnologias associadas são mais maduras, com as principais inovações já consolidadas e as patentes existentes cobrindo amplamente o campo tecnológico. (2) O baixo número de classes pode refletir setores de mercado muito específicos, nos quais a atividade inventiva está concentrada em um número restrito de instituições ou grupos de pesquisa.

Para aprofundar a análise do cenário apresentado, a [Figura 45](#) apresenta a quantidade de patentes classificadas em cada classe dentro da respectiva seção. Essa visualização permite identificar possíveis tendências e áreas de concentração dentro de cada seção.

Figura 38 – Representação do quantitativo de classificações agrupadas pela seção e a classe que compõem a primeira classificação atribuída à patente no processo de depósito.

■ Seção A ■ Seção B ■ Seção C ■ Seção D ■ Seção E ■ Seção F ■ Seção G ■ Seção H



Fonte: Elaboração do autor

A – Necessidades Humanas; B – Operações de Processamento; Transporte; C – Química e Metalurgia; D – Têxteis e Papel; E – Construções Fixas; F – Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão; G – Física; H – Eletricidade

Acesso online: <<https://public.flourish.studio/visualisation/20964094/>>

A figura apresentada exibe barras representando o volume de classificações atribuídas a cada classe dentro de suas respectivas seções, omitindo os valores quantitativos para

melhor visualização do gráfico, dada a elevada quantidade de informações representadas. Apesar dessa simplificação, é possível observar que todas as seções possuem classes que concentram um maior número de classificações.

Destaca-se a Seção A, na qual a Classe 61 corresponde a aproximadamente 65% do total de classificações atribuídas à seção. Essa classe abrange patentes relacionadas às áreas de ciência médica, veterinária e higiene, setores que refletem no cenário tecnológico e econômico do país. Salienta-se que, para obter informações detalhadas sobre a finalidade de cada classe, recomenda-se acessar o endereço eletrônico <<https://ipc.inpi.gov.br/>>.

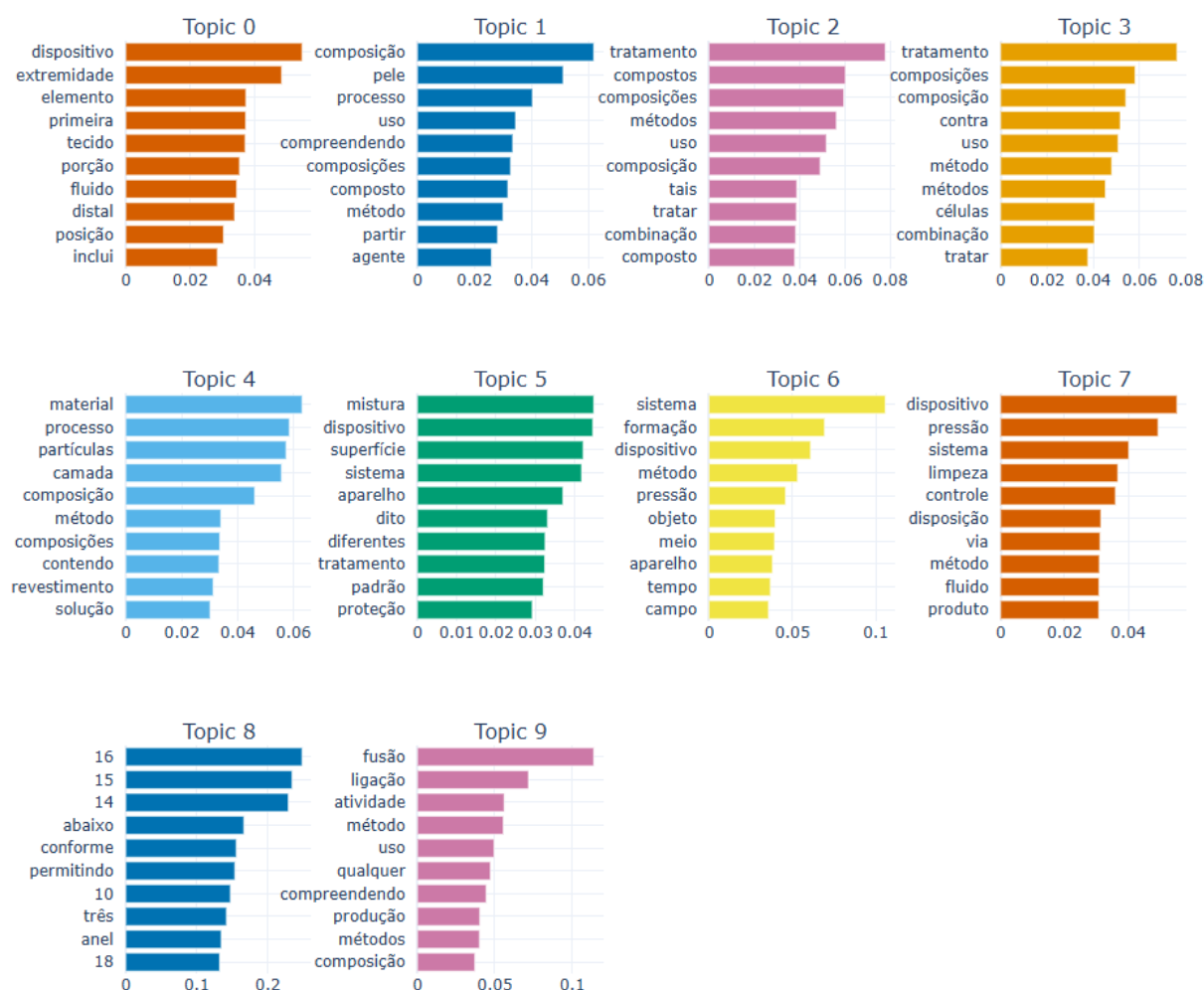
As análises realizadas até o momento foram fundamentadas em técnicas pertencentes ao campo da Patentometria, termo conceituado na [Seção 2.5](#). No entanto, os próximos resultados da pesquisa empregam metodologias baseadas em Inteligência Artificial, como por exemplo, Processamento de Linguagem Natural (PLN) e *Machine learning*. Objetivando trazer perspectivas pouco exploradas na análise de documentos de patentes.

Nesta perspectiva, a análise de frequência de termos, também conhecida como nuvem de tags, é uma técnica visual que facilita a identificação dos termos mais recorrentes nos títulos das patentes. Essa abordagem permite destacar os tópicos mais relevantes adotados nos títulos das patentes, proporcionando uma visão dos tópicos de maior relevância. Assim, a [Figura 39](#) apresenta os resultados dessa análise.



Ampliando essa análise, é possível agrupar patentes semelhantes com base nos termos utilizados em seus títulos ou resumos. Essa abordagem permite a formação de grupos de patentes que possivelmente tratam de invenções relacionadas, seja em termos de composição ou de finalidade. Também possibilita a construção de uma estrutura hierárquica, que facilita a organização e a recuperação de informações. Neste contexto, a [Figura 40](#) apresenta o resultado da análise de agrupamento de patentes, também conhecido como *clustering* ou classificação textual.

Figura 40 – Agrupamento de patentes por termos semelhantes



Fonte: Elaboração do autor

Para gerar o resultado apresentado na figura, foi selecionando um conjunto específico de patentes, foram analisados os resumos das patentes depositadas no ano de 2012 (período com o maior número de depósitos de patentes) e classificadas na Seção A e na Classe 61 (classificação com maior volume de patentes), totalizando em 4.492 patentes consideradas nesta análise. Dos 20 agrupamentos estabelecidos pela a análise, a figura apresenta os dez mais relevantes, doravante referenciados como tópicos.

Os gráficos de barras, apresentados na figura, representam os tópicos identificados, sendo que as palavras mais relevantes para cada tópico estão organizadas em ordem decrescente de importância. Cada tópico foi gerado por meio de uma metodologia que faz uso de algoritmos de processamento de linguagem natural e rede neural artificial implementados na ferramenta Pytente, que envolve a extração e agrupamento de sentenças presentes nos resumos das patentes desconsiderando as palavras sem valor semântico.

Posteriormente, foram aplicados algoritmos de *clustering* hierárquico para agrupar as patentes de acordo com a similaridade entre suas sentenças. Dessa forma, cada *cluster* formado representa um tópico distinto, caracterizado por palavras-chave extraídas dos resumos das patentes incluídas no agrupamento, agrupando as patentes com resumos semelhantes.

Essa análise possibilita a identificação dos temas mais recorrentes em um determinado conjunto de patentes, neste caso, as 4.492 patentes depositadas em 2012 e classificadas na Seção A, Classe 61. Por meio dessa abordagem, é possível inferir, por exemplo, que os tópicos 0, 1, 2 e 3 estão associados a composições, combinações e tratamentos, enquanto os tópicos 4 e 5 estão relacionados a métodos de mistura de materiais, facilitando assim a identificação de patentes de interesse.

Cada tópico é representado por um conjunto de 10 a 15 palavras-chave, que resumem os principais termos das patentes agrupadas. Essas palavras-chave, que caracterizam e distinguem cada tópico, estão detalhadas na [Tabela 12](#).

Tabela 12 – Palavras-chave que compõem os dez tópicos mais relevantes

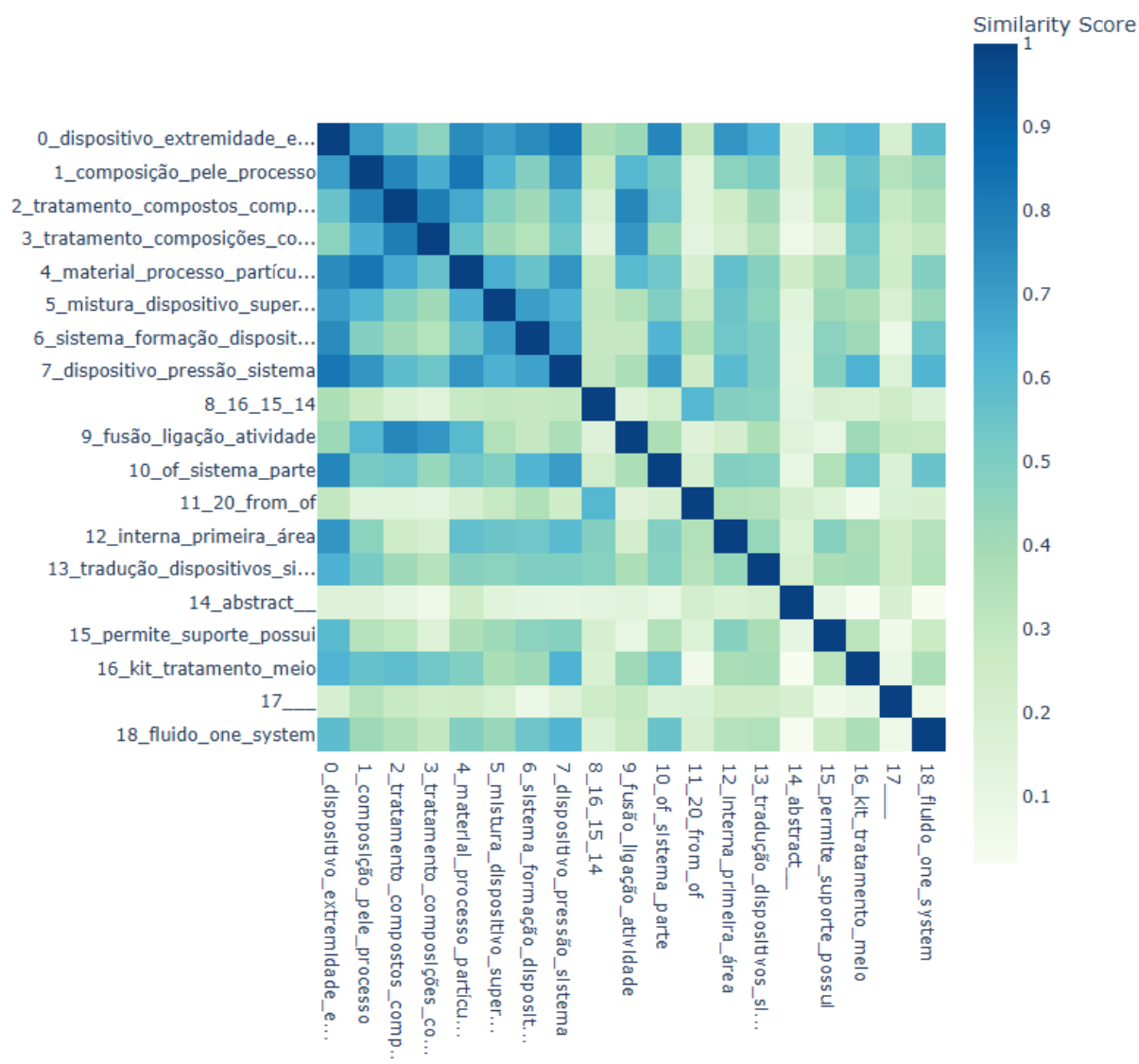
| <b>Tópico</b> | <b>Qde. Patentes</b> | <b>Palavras-chave</b>                                                                                                                                       |
|---------------|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 00            | 2.094                | dispositivo, extremidade, elemento, primeira, tecido, porção, fluido, distal, posição, inclui, primeiro, corpo, suporte, conjunto, sistema                  |
| 01            | 975                  | composição, pele, processo, uso, compreendendo, composições, composto, método, partir, agente, forma, base, sendo, tratamento, parte                        |
| 02            | 929                  | tratamento, compostos, composições, métodos, uso, composição, tais, tratar, combinação, composto, administração, indivíduo, compreendendo, método, contendo |
| 03            | 842                  | tratamento, composições, composição, contra, uso, método, métodos, células, combinação, tratar, compostos, compreendendo, atividade, ligação, tais          |
| 04            | 602                  | material, processo, partículas, camada, composição, método, composições, contendo, revestimento, solução, uso, superfície, produção, fabricação, métodos    |
| 05            | 456                  | mistura, dispositivo, superfície, sistema, aparelho, dito, diferentes, tratamento, padrão, proteção, paciente, além, composição, disso, através             |
| 06            | 350                  | sistema, formação, dispositivo, método, pressão, objeto, meio, aparelho, tempo, campo, base, dentro, paciente, alvo, 10                                     |
| 07            | 290                  | dispositivo, pressão, sistema, limpeza, controle, disposição, via, método, fluido, produto, líquido, extremidade, paciente, 10, indivíduo                   |
| 08            | 275                  | 16, 15, 14, abaixo, conforme, permitindo, 10, três, anel, 18, móvel, flexível, apresenta, interna, maior                                                    |
| 09            | 205                  | fusão, ligação, atividade, método, uso, qualquer, compreendendo, produção, métodos, composição, usos, acima, específico, composições, fornece               |
| 10            | 196                  | of, sistema, parte, and, dispositivo, através, tubular, dispositivos, superfície, liberação, passagem, interior, flexível, conjunto, método                 |

A tabela apresenta os dez tópicos mais relevantes identificados na análise. A coluna “Qtd. Patentes” exibe a quantidade de patentes associadas a cada tópico, ressaltando que uma mesma patente pode pertencer a mais de um tópico. Essa abordagem facilita o processo de pesquisa, uma vez que, ao invés de examinar todas as 4.492 patentes do conjunto, é possível concentrar a análise apenas nas patentes pertencentes aos grupos de interesse.

Conforme exposto, uma patente pode estar associada a mais de um tópico, assim como determinados termos pertencerem a diferentes tópicos, o que indica a existência de relações conceituais entre os tópicos. Por exemplo, os termos “composição” e “tratamento” aparecem em diversos tópicos, sugerindo que a elaboração de novas composições e o desenvolvimento de tratamentos constituem temas centrais neste conjunto de patentes.

Nesse contexto, a [Figura 41](#) apresenta a matriz de similaridade entre os dezenove tópicos dos vinte tópicos identificados pela análise. Essa representação facilita a visualização das similaridades existentes entre os tópicos.

Figura 41 – Matriz de similaridade entre os tópicos



Fonte: Elaboração do autor

A matriz de similaridade semântica, apresentada na figura acima, exibe as inter-relações entre 19 dos 20 agrupamentos (tópicos) estabelecidos pela análise. O vigésimo grupo foi deliberadamente excluído da visualização por constituir um agrupamento residual, composto por patentes que não puderam ser classificadas segundo as estratégias adotadas.

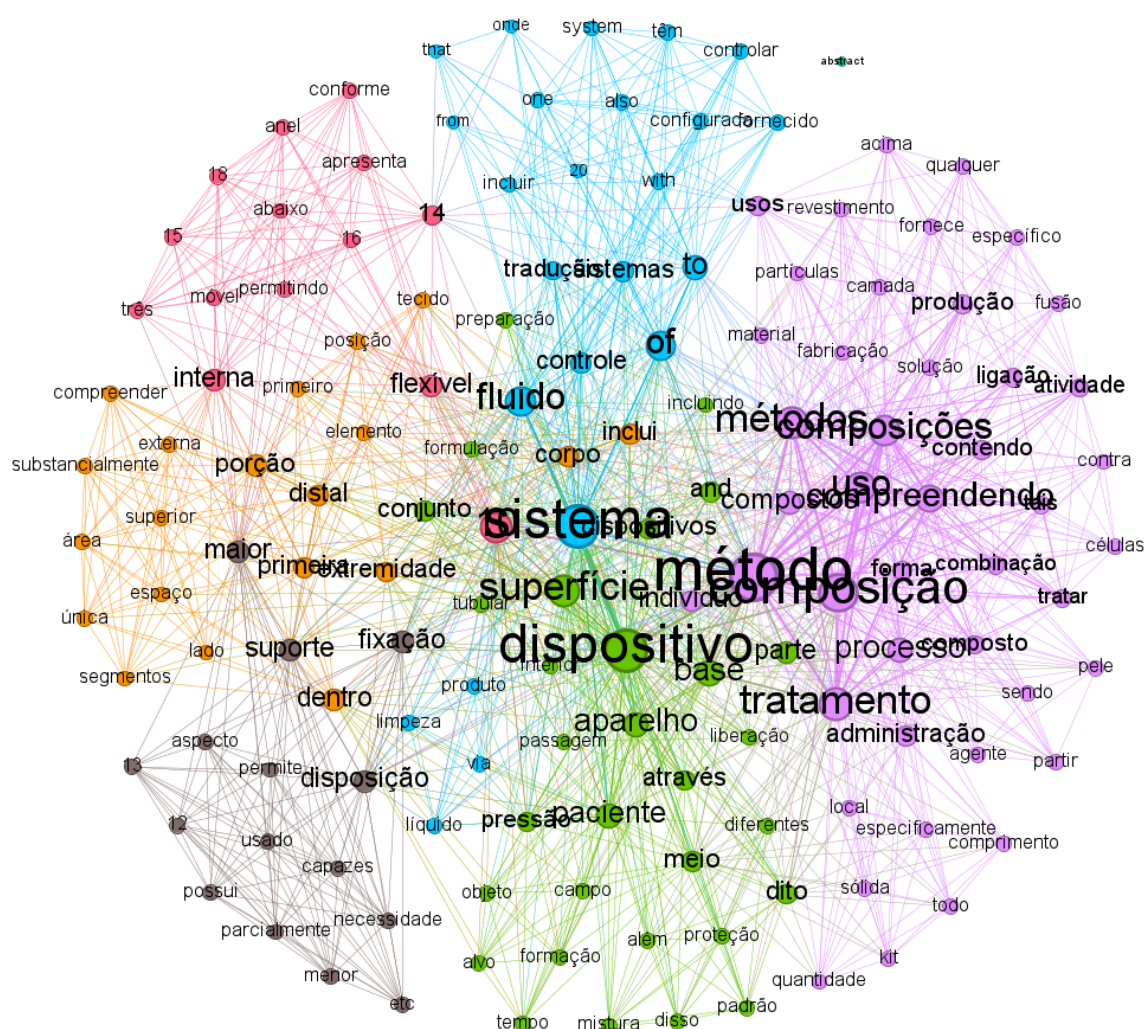
Na matriz, o grau de similaridade entre os tópicos é representado pela intensidade da cor, correspondendo ao Score de Similaridade. Valores próximos a 1 (azul intenso) denotam uma elevada proximidade semântica, enquanto tons claros indicam distanciamento terminológico.

Os resultados revelam núcleos de alta densidade correlacional, especificamente entre os tópicos 1, 2, 3 e 4, bem como no bloco formado pelos tópicos 4, 5, 6 e 7. Essa sobreposição sugere a existência de domínios tecnológicos transversais, onde as fronteiras

semânticas são fluidas. Em contraste, os tópicos 8, 11, 14 e 17 apresentam baixa similaridade com o restante do corpus. Esses agrupamentos podem ser interpretados como nichos tecnológicos isolados ou áreas emergentes que utilizam um vocabulário técnico altamente especializado e distinto, exigindo uma análise granular para identificar sua singularidade e seu potencial disruptivo no cenário da inovação.

Compreender as relações estabelecidas entre os termos contribui para identificar padrões e tendências relevantes no contexto da análise patentária. Nesse sentido, a [Figura 42](#) apresenta uma rede construída a partir dos relacionamentos entre os termos listados na [Tabela 12](#).

Figura 42 – Rede de relacionamento estabelecida entre os termos que caracterizam os principais tópicos identificando pela análise



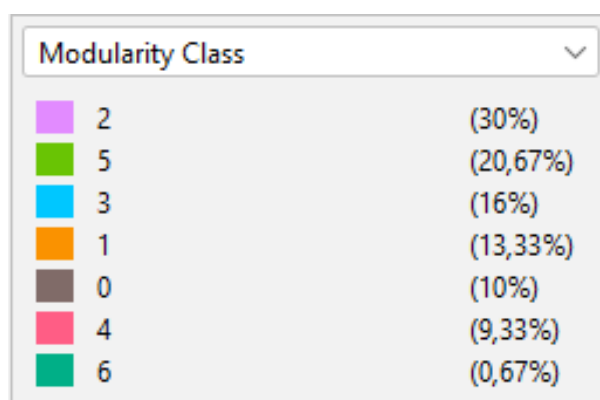
Fonte: Elaboração do autor

A rede, apresentada pela figura, exhibe os relacionamentos existentes entre os termos dos tópicos mais relevantes. A rede é composta por 150 nós, correspondendo ao número de termos distintos, e 1.568 arestas, que indicam as conexões estabelecidas entre os

termos. Entre os termos analisados, os dez com maior número de relacionamentos são: Método (88 conexões), Sistema (78 conexões), Dispositivo (76 conexões), Composição (66 conexões), Tratamento (53 conexões), Métodos (52 conexões), Superfície (51 conexões), 10 (49 conexões), Uso (47 conexões) e, por fim, Composições (47 conexões).

Em média, cada termo apresenta cerca de 20 conexões, refletindo um grau relevante de relação semântica entre os termos. Contudo, embora a rede seja formada pelos termos que compõem dez tópicos distintos, mediante a análise da rede de relacionamentos, é possível agrupar os termos em apenas sete grupos principais, representados pelas cores na visualização da rede (Figura 43). Isso evidencia a forte conexão existente entre os tópicos, que reforça a interdependência temática entre os termos analisados.

Figura 43 – Indicador de modularidade de classe da rede de relacionamento de termos



Fonte: Elaboração do autor

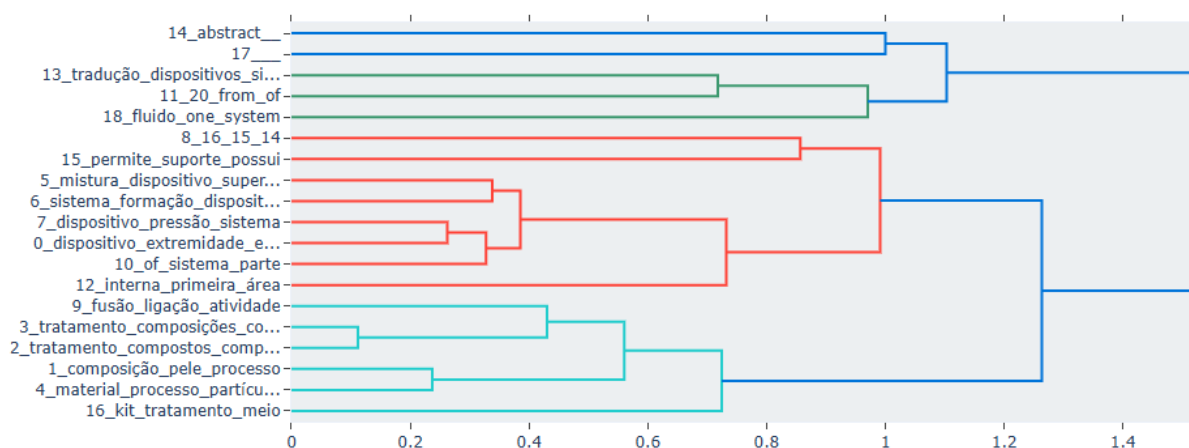
A Figura 43 apresenta as comunidades identificadas, mediante a análise de modularidade, processo que investiga a existência de estruturas modulares na rede, com propósito de avaliar nós que apresentam conexões mais densas com um grupo específico do que com os nós de outros grupos. Mediante essa estratégia, foi possível identificar sete grupos principais de nós (termos), que se organizam em *clusters* caracterizados por maior densidade de conexões internas em comparação com as conexões externas aos demais grupos.

A rede apresentada possibilita a visualização da estrutura do campo tecnológico analisado, destacando os principais termos e suas relações presentes nos resumos das patentes. Esse processo viabiliza a identificação de agrupamentos mais consistentes de documentos, contribuindo para a otimização do processo de pesquisa tecnológica e facilitando a identificação de informações relevantes.

Em um contexto hierárquico, a Figura 44 apresenta uma representação gráfica em formato de árvore, que mostra a relação hierárquica estabelecida entre os tópicos identificados. Cada folha da árvore representa um tópico específico, enquanto as linhas que

conectam as folhas indicam os graus de similaridade estabelecidos entre esses tópicos.

Figura 44 – Árvore de hierarquia de tópicos



Fonte: Elaboração do autor

A estrutura em árvore é pertinente, pois revela a hierarquia existente entre os tópicos analisados. Os tópicos mais abrangentes ocupam os níveis superiores da árvore, enquanto os tópicos mais específicos estão nos níveis inferiores, refletindo as relações de abrangência e detalhamento entre os tópicos.

A aplicação da estratégia de busca semântica demonstrou-se eficaz na recuperação de documentos relevantes a partir de descrições textuais formuladas em linguagem natural. Os resultados obtidos indicam que as respostas com índice de similaridade inferior a 30% (0.3) podem ser descartadas, por apresentarem baixo grau de relevância em relação ao termo de busca, considerados como ruído informacional.

O corpus utilizado nesta abordagem é composto por 831.706 registros de patentes únicas, totalizando aproximadamente 1 gigabyte de dados textuais. O tempo necessário para a execução completa das fases de pré-processamento e geração dos vetores de embeddings foi de aproximadamente 204 horas. Este processamento resultou na criação de 51 índices distintos, segmentados por ano de depósito das patentes, sendo realizado em um ambiente computacional com 16 GB de memória RAM e processador Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz. Importante destacar que esse processo é necessário ser executado apenas uma única vez, ou quando os dados considerados forem atualizados.

Embora envolva a varredura sequencial dos índices de embeddings, o algoritmo retorna resultados parciais em tempo real, proporcionando ao usuário a percepção de agilidade durante a consulta. Em termos de desempenho, a busca completa é concluída, em média, em menos de dois minutos, evidenciando a viabilidade da solução proposta para ambientes de pesquisa que demandam respostas rápidas e contextualizadas.

A relevância dos resultados obtidos por meio da busca semântica também se

mostrou satisfatória, uma vez que, em todos os casos analisados, as patentes retornadas apresentaram correspondência significativa com os termos expressos na consulta em linguagem natural. Essa coerência evidencia a eficácia do modelo na identificação de documentos de patentes.

A Tabela 13 apresenta uma amostragem representativa dos resultados gerados a partir da busca semântica com o seguinte enunciado: "Dispositivo ou sistema de refrigeração ou resfriamento". A seleção contempla os registros com maiores índices de similaridade.

Tabela 13 – Amostragem de resultados da busca semântica

| Relevância | Número           | Descrição resumida                                                                                                                                                                                      |
|------------|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 56,05      | BR112015017785A2 | 1/1 resumo sistema de descongelamento para aparelho de refrigeração e unidade refrigerante um sistema de descongelamento inclui: um dispositivo de esfriamento que é disposto em um freezer, e inclui   |
| 55,34      | BR0105644A       | "EQUIPAMENTO DE REFRIGERAÇÃO APERFEIÇOADO E SISTEMA DE REFRIGERAÇÃO APERFEIÇOADO". Notadamente de um equipamento e respectivo sistema, através dos quais consegue-se vantagens e melhorias significati  |
| 55,08      | BR9405086A       | Sistema de refrigeração para aparelho de refrigeração                                                                                                                                                   |
| 54,86      | BR0303842B1      | SISTEMA DE ABASTECIMENTO DE FORMAS DE GELO EM APARELHOS DE REFRIGERAÇÃO                                                                                                                                 |
| 54,48      | BRPI0614107A2    | MÉTODO E APARELHO PARA PROVER REFRIGERAÇÃO A UM DISPOSITIVO. Um sistema para esfriar um ou mais dispositivos supercondutores discretos (21,22,23) e que um refrigerador primário (1) sub-esfria liquid  |
| 54,41      | BR112015017791A2 | sistema de descongelamento por sublimação e método de descongelamento por sublimação para aparelho de refrigeração                                                                                      |
| 53,71      | BRPI0601266A     | "SISTEMA DE CONDENSACÃO APLICADO NO SISTEMA DE REFRIGERAÇÃO DE EQUIPAMENTOS DE REFRIGERAÇÃO E OU CONGELAMENTO". O presente privilégio revela um aparato que destina ser utilizado nos equipamentos de   |
| 52,85      | BR102014023977A2 | sistema conversível para congelamento e descongelamento de produtos alimentícios em aparelhos eletrodomésticos. a presente invenção pertence ao campo tecnológico dos aparelhos eletrodomésticos, e ref |
| 52,71      | BR102014005557A2 | Sistema de refrigeração de refrigerador apresentando um circuito de refrigeração secundário. A presente invenção refere-se a um sistema e método de refrigeração de refrigerador que proveem a refrige  |
| 52,31      | BR9503236A       | Sistema de degelo para aparelho de refrigeração                                                                                                                                                         |
| 52,13      | BR112015006534A2 | 1 / 1 resumo à refrigerador, e, mã todo para controlar o sistema de resfriamento de um refrigeradorã se proporciona um refrigerador (2) compreendendo um sistema de resfriamento (4) sendo atravessado  |
| 51,96      | BR112018004427B1 | DISPOSITIVO E MÉTODO DE RESFRIAMENTO PARA O RESFRIAMENTO SECUNDÁRIO DE UM LINGOTE. A invenção refere-se a um dispositivo de resfriamento (7) e a um método de resfriamento para o resfriamento secundá  |
| 51,45      | BR112015017789A2 | 1/1 resumo sistema de descongelamento para aparelho de refrigeração e unidade de resfriamento um sistema de descongelamento inclui: um dispositivo de resfriamento, que é disposto em um freezer, e in  |
| 51,24      | BRPI0815928A2    | equipamento para uso com um sistema que possui um componente de resfriamento criogênico, método utilizado com um sistema que possui um componente de resfriamento criogênico                            |
| 51,10      | BRPI0803841A2    | SISTEMA PARA MOVIMENTAÇÃO DE UM CONJUNTO DE PRATELEIRAS DE EQUIPAMENTOS DE REFRIGERAÇÃO E EQUIPAMENTO DE REFRIGERAÇÃO. A presente invenção refere-se a um sistema capaz de prover um deslocamento em t  |
| 51,09      | BR112022009060A2 | MÁQUINA DE GELO DE RESFRIAMENTO DIRETO. Um aparelho de refrigeração inclui um compartimento de alimentos frescos para armazenar itens alimentícios em um ambiente refrigerado tendo uma temperatura al  |
| 50,84      | BR112018013683A2 | trata-se de um aparelho de refrigeração que inclui um compartimento de alimento fresco e um compartimento de congelador. uma máquina de produzir gelo com uma fôrma de gelo é disposta dentro do compa  |

Fonte: Elaboração do autor

A tabela apresenta, em sua primeira coluna, o percentual de similaridade entre cada patente recuperada e o termo de busca utilizado. A segunda coluna exibe o número de publicação da patente, enquanto a terceira coluna mostra, de forma truncada, os primeiros 200 caracteres da descrição da patente tal como informada no processo de depósito. Essa limitação visa otimizar a visualização dos dados apresentados na tabela.

É possível observar que todos os 17 registros exibidos demonstram relação semântica direta com o enunciado fornecido como critério de busca. Tal fato reforça a eficácia da abordagem semântica na recuperação de documentos, uma vez que dispensa a formulação exata de palavras-chave, característica típica das abordagens tradicionais. O modelo

utilizado é capaz de reconhecer variações dos termos, como refrigeração, refrigerado, refrigerante, mantendo a coerência do significado original, o que contribui para obter resultados mais relevantes para a busca.

De maneira complementar, a [Tabela 14](#) apresenta os resultados classificados como de baixa relevância, ou seja, registros que, apesar de retornados pela busca semântica, apresentam baixa similaridade com o termo consultado, sendo considerados ruídos informacionais. Esses dados são essenciais para avaliar a precisão do modelo e delimitar o limite mínimo de similaridade aceitável para filtragem dos resultados em aplicações práticas.

Tabela 14 – Amostragem de resultados menos relevantes da busca semântica

| Relevância | Número     | Descrição resumida                                                                                                                                          |
|------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 27,88      | BR8007413A | SISTEMA DE MANIPULACAO DE FILEIRA E DISPOSITIVO PRENDEDOR DE FILEIRA                                                                                        |
| 27,74      | BR8006433A | PROCESSO, APARELHO, SISTEMA E DISPOSITIVO PARA REFINACAO DE MATERIAIS FIBROSOS                                                                              |
| 27,19      | BR7407402A | SISTEMA DE SUSPENSÃO PARA UMA UNIDADE ATRELADA                                                                                                              |
| 27,12      | BR7409009A | PROCESSO E DISPOSITIVO PARA TROCA DE MATERIAL ENTRE SISTEMAS HETEROGENEOS                                                                                   |
| 26,34      | BR7402474A | DISPOSITIVO E SISTEMA PARA DETETAR PARTICULAS SOLIDAS NUM FLUXO DE CORRENTE                                                                                 |
| 26,27      | BR7403478A | SISTEMA DE COLOCACAO DE APARELHO ECONOMIZADOR DE GASOLINA                                                                                                   |
| 26,00      | BR7409067A | SISTEMA DE CONTROLE PARA EQUIPAMENTO DE MOLDAGEM DE SOLADOS PARA CALCADOS                                                                                   |
| 25,70      | BR7406594A | APERFEICOAMENTO EM APARELHO DE EVAPORACAO PARA USO NA CONCENTRACAO POR EVAPORACAO E/OU PURIFICACAO DE SOLUCOES E SISTEMA DE EVAPORACAO DE MULTIPLOS EFEITOS |
| 25,66      | BR7403450A | DISPOSITIVO E PROCESSO DE SUPRIMENTO DE FLUIDO A UM SISTEMA INFLAVEL PARA AMPARAR O OCUPANTE DE UM VEICULO                                                  |
| 24,34      | BR7405136A | PROCESSO E APARELHO PARA RESFRIAR UM MATERIAL QUENTE EM PARTICULAS                                                                                          |
| 23,13      | BR7402628A | SISTEMA DE FRITAR COM AGUA E AZEITE A TEMPERATURA VARIÁVEL                                                                                                  |
| 23,12      | BR5401564U | PROCESSO DE INSTALACAO E ADAPTACAO DE FECHADURAS PARA GELADEIRAS                                                                                            |
| 21,99      | BR7403207A | MAQUINA ENSACADEIRA DE CARNES PARA OBTENCAO DE FRIOS EM GERAL                                                                                               |
| 21,83      | BR5401565U | PROCESSO DE INSTALACAO E ADAPTACAO DE FECHADURAS PARA GELADEIRAS                                                                                            |
| 21,38      | BR7403284A | DISPOSITIVO PARA A VENTILACAO E EXAUSTAO DE AMBIENTES EDIFICIOS E SIMILARES E DISPOSITIVO PARA A PERMUTA DE CALOR DE DOIS CIRCUITOS DE ELEMENTOS            |
| 21,31      | BR7410986A | DISPOSITIVO DE BOMBA PARA CONTROLE DE PROPORCAO REGIME VELOCIDADE DA CIRCULACAO DE LIQUIDO NUM SISTEMA DE ENCANAMENTO                                       |
| 21,30      | BR5403295U | CAIXA AUTO-ESTRUTURADA PARA REFRIGERADOR                                                                                                                    |
| 21,13      | BR7408722A | SISTEMA APERFEICADO DE AUMENTO DE INTUICAO DE BOAS CONDICAOES DE VOO SEM CONSULTA A APARELHO DE MEDIDA                                                      |
| 21,03      | BR7402890A | SISTEMA DE APLICACAO DE UNIDADES COM CARACTERISTICAS PROPRIAS DE PIEZOELETRICO EM FOGOS FOGAREIROS                                                          |
| 20,57      | BR5403359U | AQUECEDORES FORNOS E SEUS SIMILARES                                                                                                                         |
| 19,43      | BR7506158A | CAIXA AUTO-ESTRUTURADA PARA CONGELADORES                                                                                                                    |
|            |            | SISTEMA DE PROTENSÃO                                                                                                                                        |

Fonte: Elaboração do autor

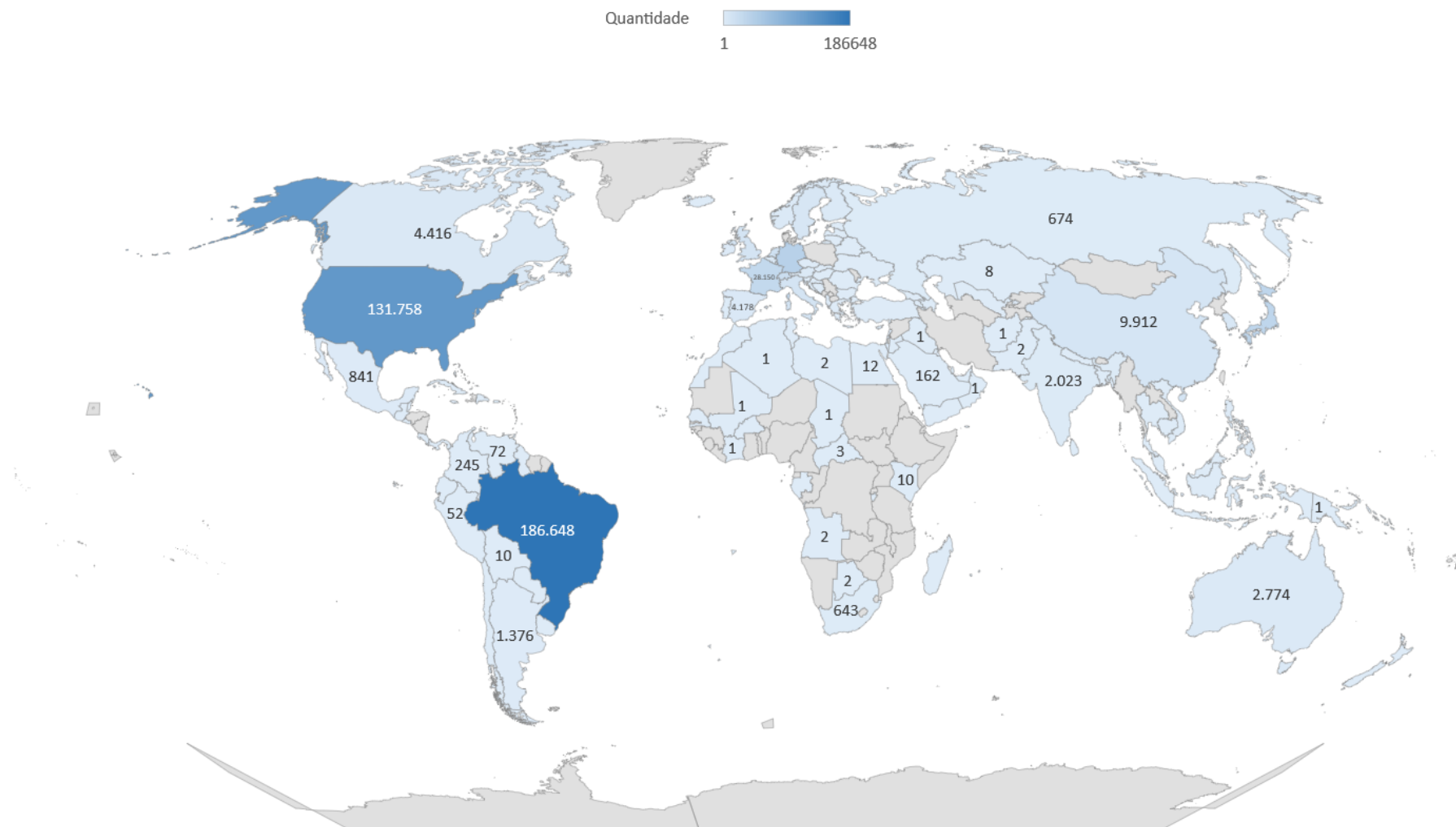
Conforme evidenciado na [Tabela 14](#), os resultados que apresentam percentual de similaridade inferior a 30% demonstram, de fato, baixa relevância em relação ao termo de busca utilizado. Esses registros incluem patentes que não abordam diretamente tecnologias ou dispositivos relacionados a sistemas de refrigeração ou resfriamento, sendo considerados ruído informacional no contexto da consulta semântica. Esses resultados reforçam a estratégia adotada neste estudo, na qual apenas patentes com similaridade superior a 30% devem ser consideradas para fins analíticos, a fim de garantir maior precisão na recuperação da informação e relevância dos resultados obtidos.

Em sequência, mediante a adoção de fontes complementares, é possível desenvolver análises que caracterizam as pessoas envolvidas no processo de depósito. Abordagens que

não seriam possíveis utilizando apenas o conjunto de dados principal, oriundo da Espacenet. Um exemplo disso é a análise do perfil dos depositantes, que pode ser realizada com base no país de residência dos depositantes, no caso de pessoas físicas, ou no país de sede da organização, quando o depositante for uma pessoa jurídica.

Devido à inexistência de uma patente com validade mundial, a proteção conferida a um invento é limitada ao território do país onde o depósito foi realizado. Consequentemente, para assegurar proteção em outros países, é necessário efetuar depósitos adicionais, formando assim uma família de patentes. Nesse contexto, a [Figura 45](#) apresenta a distribuição total de depósitos de patentes segundo o país de residência dos depositantes. Essa visualização foi gerada com base em um conjunto de dados complementar oriundo do INPI, abrangendo informações de 154.291 depositantes.

Figura 45 – Total de depositantes por nacionalidade informada no processo de depósito de patente



Fonte: Elaboração do autor

Os Estados Unidos se destacam como o principal país de origem dos depositantes, com um número expressivo de patentes depositadas no Brasil. Essa concentração indica a presença expressiva de empresas e produtos norte-americanos no mercado brasileiro, seja por meio de filiais locais ou parcerias estabelecidas com empresas nacionais.

Em menor escala, outros países, como China, Japão e países europeus, também desempenham um papel relevante no depósito de patentes no Brasil. Essa participação evidencia o interesse estratégico desses países em proteger suas inovações no território brasileiro. Vale ressaltar que a análise foi baseada na nacionalidade do depositante informado durante o processo de depósito, o que, em alguns casos, pode não corresponder ao país de origem da patente.

A ampla distribuição geográfica dos depositantes, apresentada pela figura, reflete o impacto da crescente globalização da inovação. Empresas de diferentes partes do mundo estão buscando assegurar proteção a suas invenções em mercados estratégicos, como o brasileiro, o que reforça a relevância do Brasil como um importante polo para a proteção de inovações no cenário global.

Conforme apresentado, esta seção foi dedicada à análise exploratória da base de dados local, construída de maneira sistemática em conformidade com as estratégias delineadas neste estudo. Foi possível apresentar um panorama abrangente da expressiva produção técnica brasileira, registrada tanto em documentos de patentes.

Os resultados obtidos, inclusive aqueles derivados de técnicas avançadas como Processamento de Linguagem Natural (PLN) e *Machine Learning*, podem ser facilmente aplicados a diferentes conjuntos de dados. Para isso, basta ajustar os parâmetros de entrada da ferramenta Pytente. A ferramenta Pytente oferece um ambiente intuitivo e acessível, especialmente projetado para atender às necessidades de pesquisadores com pouca ou nenhuma familiaridade com recursos computacionais, como linguagens de programação ou algoritmos de *Machine Learning*.

A próxima seção é dedicada à qualificação e certificação dos dados de patentes coletados nos currículos da Plataforma Lattes.

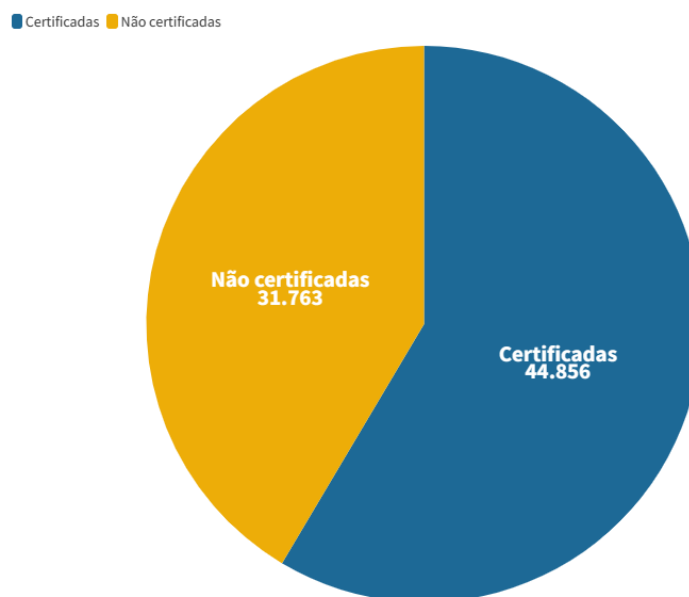
### 7.3 Certificação de patentes na Plataforma Lattes

Considerando a produção técnica brasileira, registrada na Plataforma Lattes, foram recuperados 76.619 registros de patentes e 19.264 registros de programas de computador com registro. Por meio das estratégias metodológicas propostas neste estudo, foi possível certificar 44.856 registros de patentes. Em outras palavras, das 76.619 patentes informadas pelos proponentes em seus currículos, apenas 44.856 puderam ser identificadas no conjunto de patentes proveniente do repositório Espacenet, o que corresponde a aproximadamente

58,5% do total de registros do conjunto PGPL.

A Figura 46 apresenta, em um gráfico comparativo, a distribuição entre patentes certificadas e não certificadas, permitindo uma análise clara da proporção de registros efetivamente validados em relação ao total declarado.

Figura 46 – Patentes certificadas e patentes não certificadas informadas no currículos da Plataforma Lattes



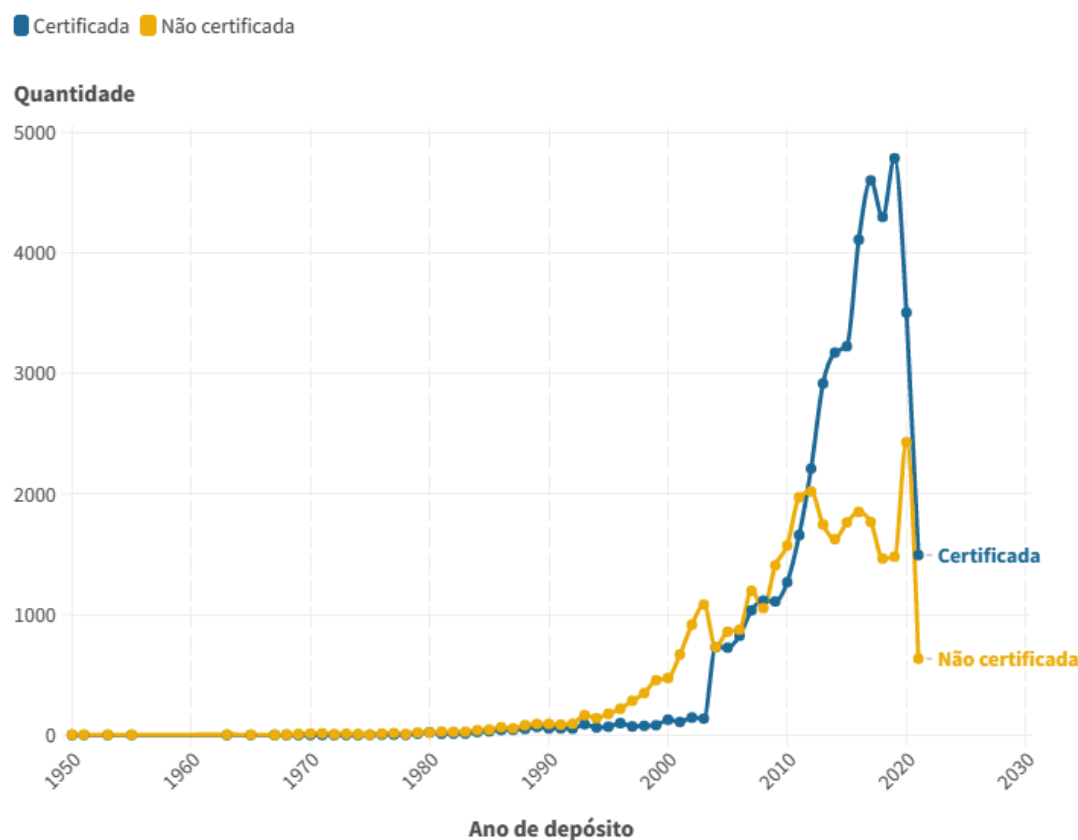
Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/13572201/>>

É possível observar que pouco mais da metade das patentes declaradas nos currículos registrados na Plataforma Lattes foram certificadas por meio das estratégias metodológicas desenvolvidas neste estudo. Esse resultado destaca a importância de implementar mecanismos de validação para as informações fornecidas pelos proponentes em seus currículos, uma vez que o processo de inserção manual está sujeito a erros e inconsistências.

Com o objetivo de aprofundar essa análise, a Figura 47 apresenta a distribuição temporal do total de depósitos de patentes por ano, considerando as informações registradas nos currículos. O gráfico diferencia duas séries: uma representando as patentes certificadas e a outra correspondente às não certificadas. Essa visualização permite uma avaliação das discrepâncias existentes entre os dois grupos.

Figura 47 – Patentes certificadas e patentes não certificadas por ano de depósito informado nos currículos da Plataforma Lattes



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/25974071/>>

É possível observar que a patente mais antiga registrada nos currículos da Plataforma Lattes foi depositada no ano de 1950. Entre os anos de 1950 e 1990, o número de patentes certificadas e não certificadas permaneceu relativamente próximo, uma vez que o volume total de registros nesse período era baixo. No entanto, entre 1990 e 2002, houve um aumento significativo no número de patentes informadas nos currículos, acompanhado por um crescimento expressivo no número de patentes não certificadas. É importante destacar que esta análise considerou somente os registros que possuíam o ano de depósito informado.

Uma hipótese para esse aumento das patentes não certificadas está relacionada às mudanças nos formatos de numeração das patentes adotadas pelo INPI. Inicialmente, o INPI utilizava um formato específico para a numeração de depósitos de patentes. A partir de 2 de janeiro de 2012, esse formato foi alterado para atender ao padrão internacional de identificação de patentes, padrão também adotado pela base Espacenet. Essa mudança implicou que patentes depositadas no Brasil até o ano de 2011 tivessem sua numeração

adaptada, de forma a alinhar com o padrão internacional.

Essas alterações podem explicar o aumento no número de patentes não certificadas entre 1990 e 2002, já que muitas delas provavelmente foram informadas nos currículos no formato antigo e, portanto, não puderam ser associadas diretamente às informações disponíveis na Espacenet. Após 2012, com a adoção do padrão internacional pelo INPI, o número de patentes certificadas superou o de não certificadas. Esse resultado corrobora a hipótese de que as patentes anteriores a 2012 não passaram pelo processo de certificação devido ao formato do número da patente ser diferente do esperado pela Espacenet.

Entre 1950 e 2011, foram certificadas 10.661 patentes, representando 41,86% dos registros informados nesse período, enquanto 14.803 registros (58,13%) não foram certificados. Já no intervalo de 2012 a 2021, é possível observar uma mudança significativa: 34.195 registros foram certificados, correspondendo a 66,85%, enquanto 16.950 (33,14%) permaneceram não certificados.

A estratégia proposta neste trabalho certifica apenas patentes depositadas no Brasil, portanto, é importante verificar se existem patentes que não foram depositadas no Brasil, no conjunto das patentes não certificadas. Logo, foi possível identificar que 7.963 registros de patentes não certificadas não foram depositados no Brasil, justificando a sua não certificação mediante as estratégias definidas.

Após a exclusão das patentes não depositadas no Brasil, o conjunto de registros não certificados passou a totalizar 23.790 patentes, o que equivale a aproximadamente 31% das patentes informadas nos currículos registrados na Plataforma Lattes. Esses registros não puderam ser certificados mediante as estratégias adotadas neste estudo.

Ao investigar as razões para a não certificação desses registros de patentes, foram identificadas diversas inconsistências nos números de depósito fornecidos pelos proponentes. Entre os problemas encontrados, estão os números de depósito em formatos inválidos, como “62/1781”, “N.Ref. 05385”, “PI ? 0006594-3”, “n.05059911-9” e “INPI PI0705367”, e a presença de registros informados como depósitos realizados no Brasil, mas que, na realidade, não correspondem a tal condição, uma vez que não iniciam com o prefixo “BR”, como os números “WO2010054452” e “US2009246833-A1”.

Essas inconsistências evidenciam a necessidade de mecanismos de validação das informações inseridas pelos proponentes nos currículos da Plataforma Lattes.

Com intuito de verificar a eficiência da aplicação de cada regra estabelecida para o processo de certificação, conforme apresentado na [Seção 5.4](#), a [Tabela 15](#) apresenta o quantitativo de patentes identificadas mediante a aplicação de cada regra.

Tabela 15 – Detalhamento da certificação

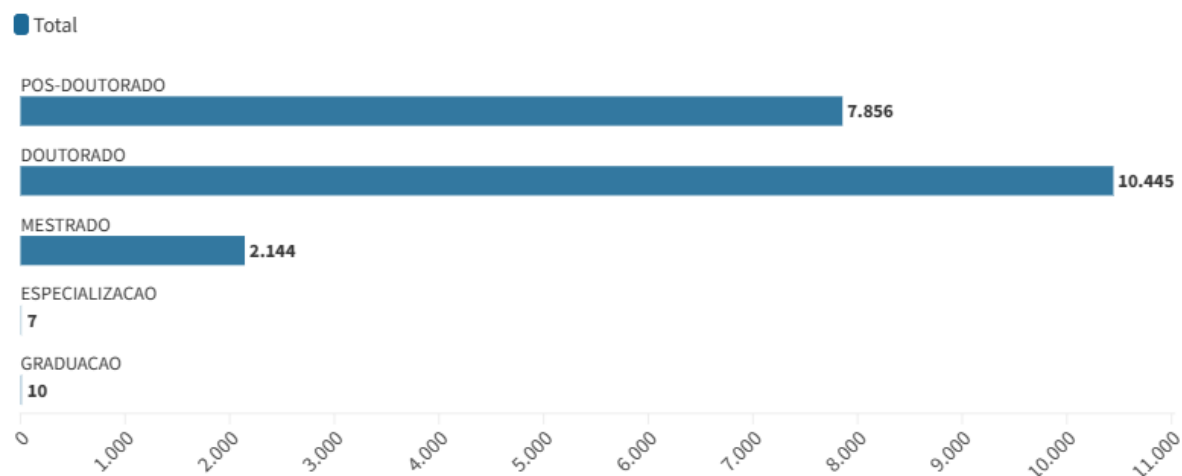
| Regra     | Quantidade |
|-----------|------------|
| Regra 001 | 1.441      |
| Regra 002 | 34.447     |
| Regra 003 | 628        |
| Regra 004 | 4.901      |
| Regra 005 | 3.098      |
| Regra 006 | 264        |
| Regra 007 | 77         |

Conforme apresentado, considerando todo o conjunto de 44.348 registros patentes certificadas, apenas 1.441 puderam ser certificadas sem nenhum tipo de tratamento específico, ou seja, foram identificados da forma que foram informados pelos seus proponentes nos currículos, aproximadamente 3,3% do conjunto de patentes certificadas.

Contudo, certificar os dados contidos na Plataforma Lattes se torna relevante, pois vários trabalhos têm se utilizado dos dados registrados em nos currículos para diversas análises ([Mena-Chalco e Cesar-Junior \(2009\)](#); [Digiampietri \(2015\)](#); [Brito, Quoniam e Mena-Chalco \(2016\)](#); [Dias, Moreira e Dias \(2018\)](#)).

Com objetivo de caracterizar os proponentes da produção técnica brasileira registrada na Plataforma Lattes, diversas análises foram realizadas. Na Plataforma Lattes é possível consultar a formação acadêmica dos proponentes. Apesar do nível de formação não ser o único indicador do potencial de um indivíduo, pode fornecer uma série de informações importantes sobre ele, suas habilidades, conhecimentos específicos e experiências de pesquisa e desenvolvimento. Um nível mais elevado de formação muitas vezes está associado a uma capacidade mais desenvolvida de análise crítica, interpretação de informações complexas e, em muitos casos, está associado a uma maior experiência em pesquisa acadêmica e desenvolvimento de projetos. Assim, a [Figura 48](#) apresenta a distribuição dos pesquisadores que possuem um ou mais registros de patentes certificadas considerando seu maior nível de formação acadêmica informado.

Figura 48 – Formação dos proponentes de patentes



Fonte: Elaboração do autor

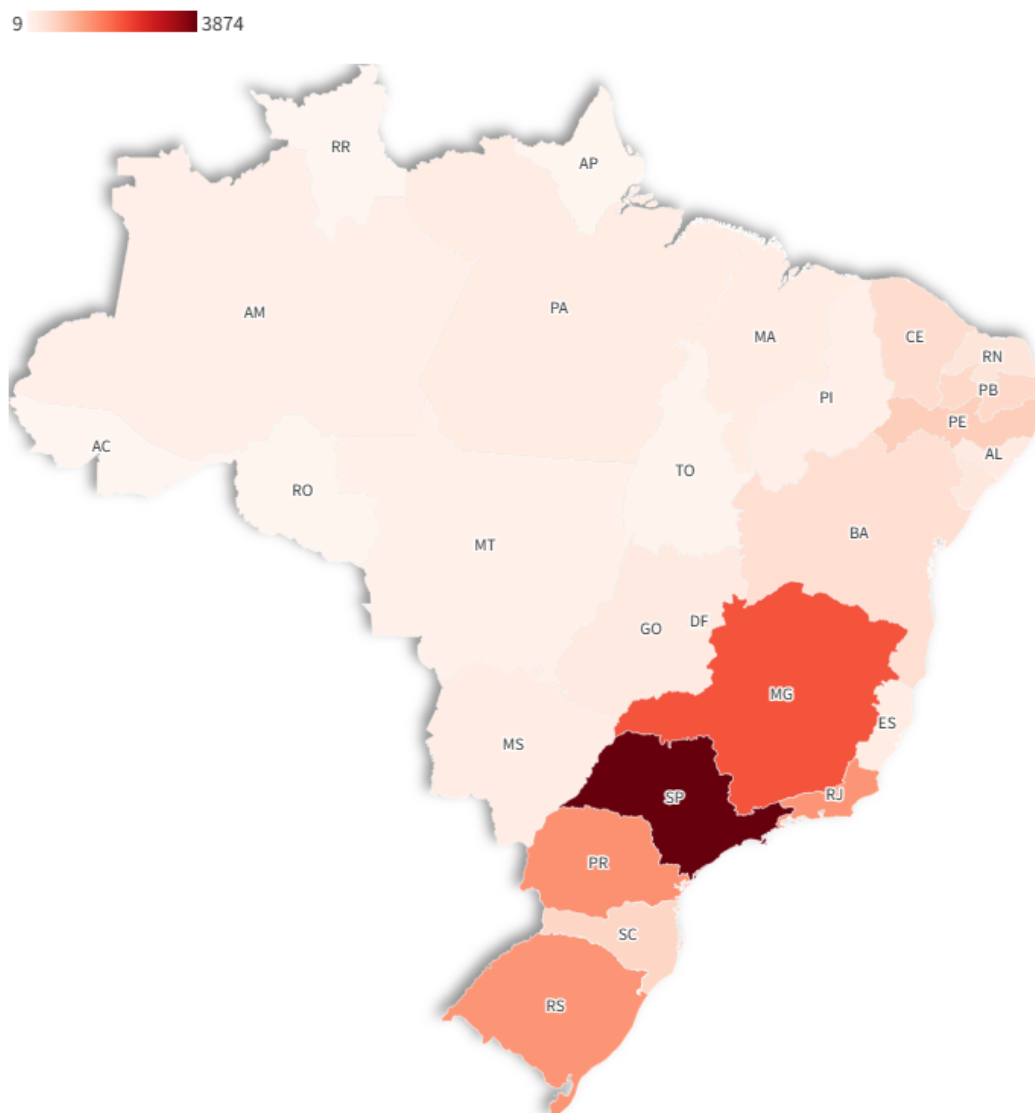
Acesso online: <<https://public.flourish.studio/visualisation/13887148/>>

Avaliar o perfil dos depositantes de patente é importante para identificar especialistas em determinadas áreas do conhecimento, bem como assinalar possíveis parcerias para o desenvolvimento de novas tecnologias. O nível de formação muitas vezes reflete o compromisso de um indivíduo com o aprendizado contínuo e o desenvolvimento pessoal, sugerindo uma disposição para se dedicar a pesquisas e desenvolvimentos. De acordo com Dias (2016), os doutores são responsáveis pela maioria das produções científicas registradas nos currículos da Plataforma Lattes, e esse fenômeno se repete na produção técnica, doutores e pós-doutores juntos correspondem a 89,4% do total de depositantes de patentes certificadas. Apesar de o pós-doutorado não ser um nível de formação acadêmica, os indivíduos foram classificados nesta categoria para melhor visualização e classificação.

Em sequência, no processo de caracterização dos indivíduos, nos currículos da Plataforma Lattes é possível informar quais as suas principais áreas de atuação. Este fato permite averiguar qual a área do conhecimento mais representativa, tendo em vista o perfil dos proponentes analisados e, consequentemente, avaliar qual área do conhecimento se caracteriza como propulsora do desenvolvimento tecnológico nacional. Na Figura 49 é possível verificar a distribuição dos proponentes de patentes agrupados por áreas de atuação. O critério de seleção adotado foi contabilizar todos os currículos que tiverem pelo menos uma patente certificada e que possuam informação de área de atuação registrada. Como um currículo pode ter até seis registros de áreas de atuação, para esta análise, a primeira área informada foi considerada.



Figura 50 – Distribuição de proponentes de patentes por estado



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/14694155/>>

As regiões mais expressivas são a Sudeste e Sul no contexto da atividade de pesquisa e desenvolvimento tecnológico no Brasil. Os estados, São Paulo (com 3.874 pesquisadores), Minas Gerais (2.183 pesquisadores) e Paraná (1.474 pesquisadores) apresentam os maiores quantitativos de indivíduos atuando nesse campo. Uma possível explicação para essa concentração está na presença das maiores universidades depositantes de patentes nessas regiões, contribuindo significativamente para o aumento do número de pesquisadores e o fortalecimento das atividades de pesquisa e inovação.

Essa concentração evidencia a posição de destaque dos estados de São Paulo, Minas Gerais e Paraná no cenário nacional de pesquisa. Esse tipo de análise é importante

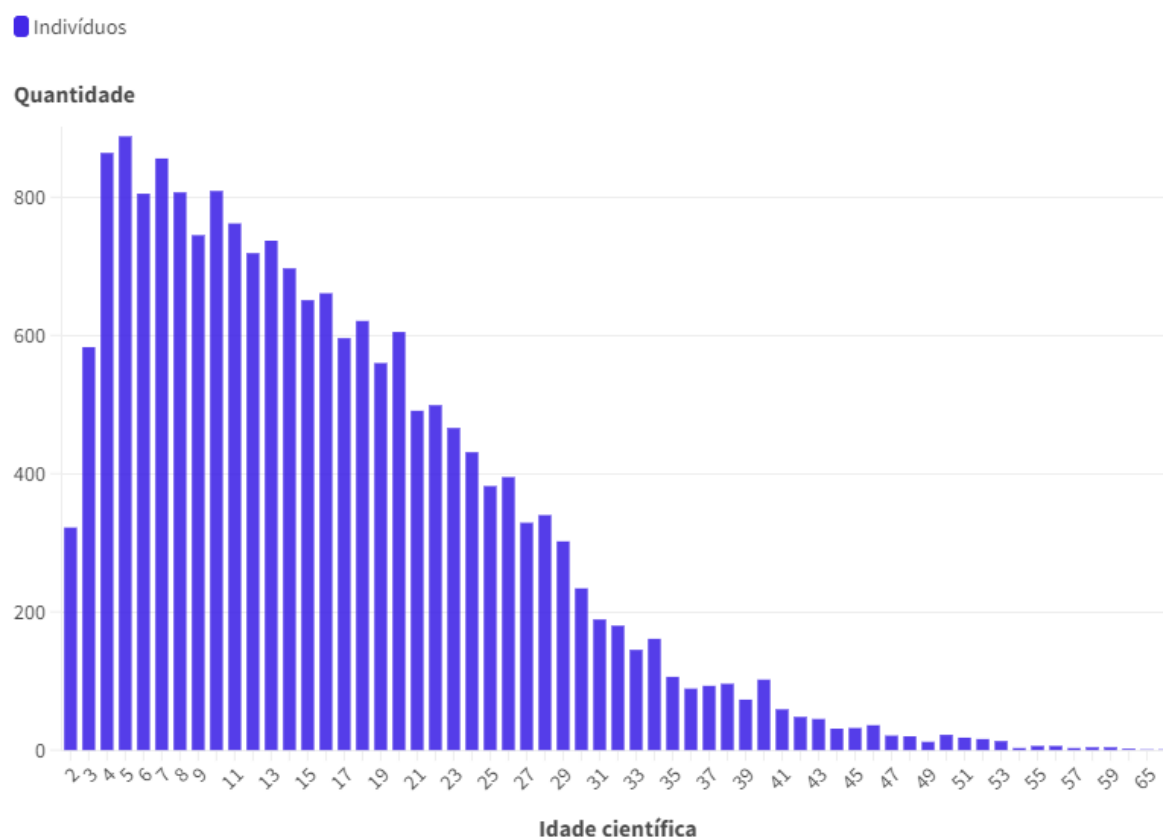
para compreender as disparidades regionais em termos de inovação e desenvolvimento científico no Brasil, além de identificar áreas que demandam maior investimento e esforços de desenvolvimento, com o objetivo de promover um equilíbrio no panorama nacional de pesquisa e inovação.

Um fator relevante para caracterizar o perfil dos proponentes de patentes registrados na Plataforma Lattes é a análise da idade científica, que consiste no cálculo do tempo decorrido desde a conclusão do doutorado do indivíduo. Nesta análise, é considerada somente a titulação de doutor, pois este grupo é amplamente reconhecido como o mais produtivo em termos de produção científica e tecnológica.

Muitos desses doutores estão vinculados a programas de pós-graduação ou são beneficiários de bolsas de produtividade em pesquisa concedidas pelo CNPq, o que os incentiva a manter seus currículos constantemente atualizados. Além disso, a maioria desses profissionais está associada a grupos de pesquisa registrados no Diretório de Grupos de Pesquisa do CNPq, reforçando sua integração com o ambiente acadêmico e científico no Brasil (DIAS, 2016).

Nesse contexto, a [Figura 51](#) apresenta a distribuição da idade científica dos proponentes cujas patentes registradas na Plataforma Lattes foram certificadas.

Figura 51 – Idade científica

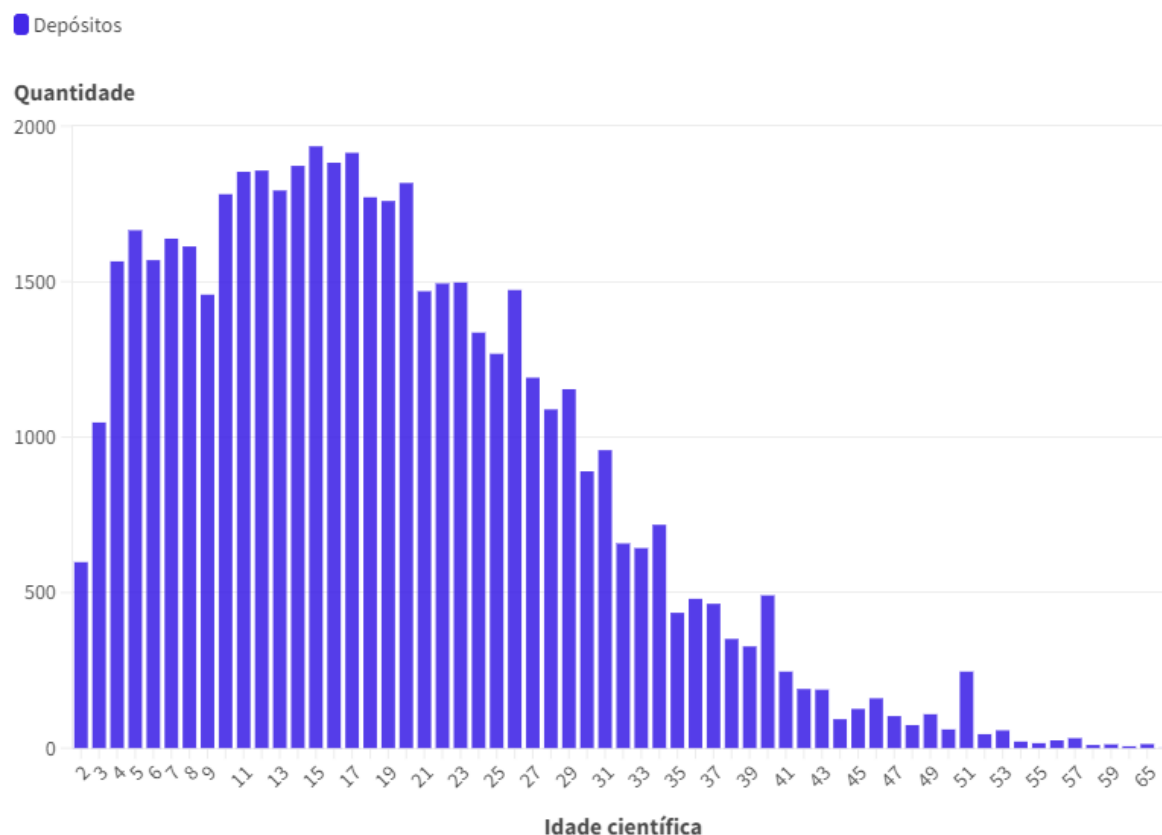


Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/14903395/>>

De acordo com a figura apresentada, a média de idade científica é de 32 anos, sendo que a maioria dos doutores (888 indivíduos) têm 5 anos de obtenção do título. Aproximadamente metade do conjunto considerado na análise (9.594) possui idade científica entre 2 e 14 anos. Contudo, é importante ressaltar que o tempo de conclusão do doutorado não está diretamente ligado ao número de depósito de patentes, embora a experiência possa contribuir significativamente para a produtividade, a qualidade e a quantidade do trabalho produzido por um pesquisador dependem de vários outros fatores. Aprofundando nesta perspectiva, a [Figura 52](#) apresenta a quantidade de patentes depositadas de acordo com a idade científica dos respectivos proponentes.

Figura 52 – Total de depósito por idade científica



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/14904090/>>

Conforme apresentado, aproximadamente metade das patentes (26.040), foram depositadas por pesquisadores com idades científica entre 2 e 17 anos, que compreende o intervalo com maior número de pesquisadores apresentado anteriormente. Portanto, embora a conclusão oportuna do doutorado possa refletir uma certa eficiência ou alta capacidade de pesquisa, isso por si só não é indicativo da produtividade do pesquisador. É importante destacar que a produtividade acadêmica é melhor avaliada por meio de métricas como publicações em revistas revisadas, citações, reconhecimento na comunidade acadêmica, contribuições significativas para o avanço de um determinado campo de estudo específico, entre outros indicadores. No entanto, através do indicador apresentado, é possível observar que o grupo de depositantes pode ser caracterizado como um grupo maduro mediante o tempo de conclusão do doutorado.

No contexto da contribuição de pesquisadores e universidades para o desenvolvimento tecnológico nacional, dados do Instituto Nacional da Propriedade Industrial (INPI) indicam que, em 2024, a Universidade Federal de Campina Grande ocupou a terceira posição no ranking de depositantes de patentes residentes no Brasil, com 86 depósitos,

ficando atrás apenas da Stellantis Automóveis Brasil, com 185, e da Petrobras, com 155 depósitos. O “Ranking de Depositantes Residentes de Patentes de Invenção – 2024”, elaborado pelo INPI, inclui 50 depositantes, dos quais 30 são universidades estaduais e federais, demonstrando a predominância dessas instituições entre os solicitantes de patentes no país (INPI, 2024b).

Inclusive, desde 2018, as instituições de ensino superior têm se destacado no cenário nacional. Em 2018, a Universidade Federal da Paraíba foi o maior depositante de patentes residentes no Brasil, com 94 depósitos; em 2019, manteve essa posição com 100 depósitos; e, em 2020, a Universidade Federal de Campina Grande, também da Paraíba, assumiu o primeiro lugar, com 96 depósitos (INPI, 2022).

Neste sentido a [Figura 53](#) apresenta 30 Instituições de ensino superior (IES) com o maior número de patentes depositadas. Para gerar a visualização foi considerado somente os depositantes que contenham os seguintes termos em seu nome: “UNIVERSIDADE” ou “FACULDADE” ou “ESCOLA” ou “CENTRO DE EDUCAÇÃO” ou “ENSINO SUPERIOR” ou “CENTRO UNIVERSITÁRIO” ou “INSTITUTO FEDERAL” ou “CENTRO FEDERAL DE EDUCAÇÃO” de todo o conjunto de patentes únicas (PEU), ou seja, essa visualização não considera somente as patentes certificadas.

Figura 53 – 30 IES depositantes de patentes



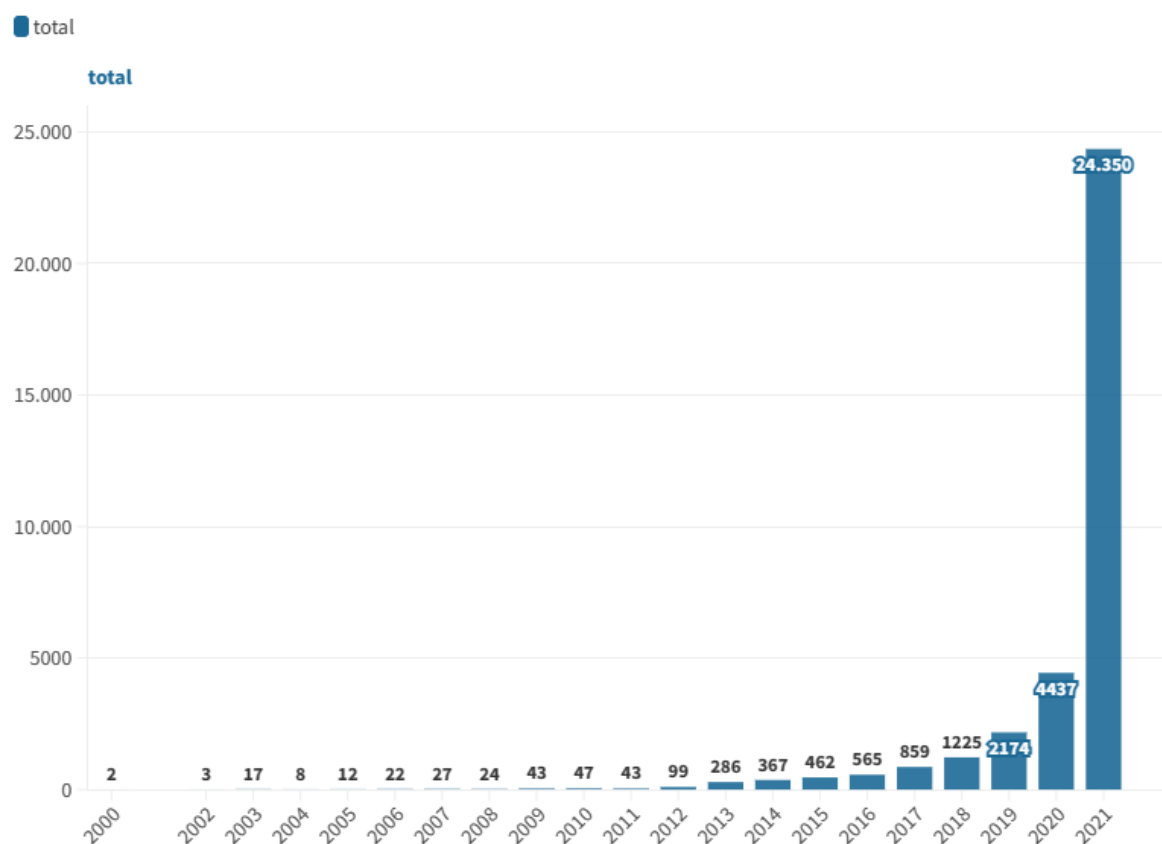
Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/14877469/>>

Conforme apresentado, a Universidade Estadual de Campinas - Unicamp (8,22%), a Universidade de São Paulo - USP (6,84%) e a Universidade Federal de Minas Gerais - UFMG (6,06%) se destacam no depósito de patentes. Todas essas universidades têm desempenhado um papel crucial na formação de profissionais altamente qualificados, na promoção da pesquisa e da inovação e no avanço do conhecimento em diversas áreas. Destacando que são universidades localizadas na região sudeste do país, reafirmando o indicador apresentado pela Figura 50.

A fim de garantir a confiabilidade nas informações apresentadas, para averiguar se as informações contidas nos currículos podem ser consideradas atualizadas, a Figura 54 apresenta o total de currículos atualizados por ano, considerando todo o conjunto de currículos coletados na Plataforma Lattes. Usando como critério o campo “Última atualização do currículo” informado nos currículos.

Figura 54 – Atualização dos currículos da Plataforma Lattes



Fonte: Elaboração do autor

Acesso online: <<https://public.flourish.studio/visualisation/13887055/>>

Mediante ao exposto, considerando a relevância dos dados contidos nos currículos da Plataforma Lattes, é essencial ressaltar que, de um total de 35.072 currículos coletados, uma expressiva maioria de 30.961 currículos revelou-se atualizada nos últimos cinco anos. Esse dado representa um percentual de 88,3% do conjunto de dados investigados. Este resultado evidencia a conscientização e o comprometimento dos profissionais em manter seus currículos devidamente atualizados. Entretanto, é importante destacar que muitas instituições acadêmicas e órgãos de fomento à pesquisa utilizam o currículo Lattes como referência para avaliação de pesquisadores e concessão de bolsas e financiamentos. Portanto, um currículo atualizado e detalhado é crucial para garantir o reconhecimento institucional e a obtenção de apoio financeiro para projetos de pesquisa.

Os indicadores gerados com base nos dados extraídos dos currículos cadastrados na Plataforma Lattes destacam que a plataforma, além de informações científicas, informações tecnológicas também podem ser recuperadas e utilizadas como subsídios para diversas análises.

Neste contexto, é importante destacar que vários registros não puderam ser anali-

sados, devido a falta de certificação mediante as estratégias propostas neste estudo. Isso reforça a importância em se utilizar de estratégias para a vinculação e posterior certificação. Tal processo se torna importante tendo em vista que viabiliza de forma efetiva a possibilidade de validar os dados dos currículos com informações complementares validadas.

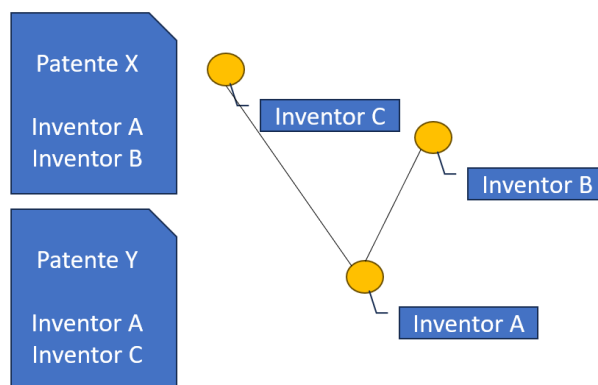
Compreender o perfil do proponente permite retratar de maneira mais precisa a credibilidade e a experiência do inventor. Isso pode contribuir para determinar a capacidade do proponente de conduzir pesquisas relevantes, propor inovações e até mesmo identificar especialistas em determinadas áreas do conhecimento. Ter conhecimento do perfil do proponente contribui na formação de colaborações e parcerias estratégicas com outras entidades interessadas no desenvolvimento tecnológico.

## 7.4 Rede de colaboração estabelecida entre os pesquisadores com currículo registrados na Plataforma Lattes

A análise de redes de colaboração científica desempenha um papel fundamental na compreensão da dinâmica e do funcionamento da comunidade científica. Essa abordagem visa mapear e analisar as relações de colaboração entre pesquisadores, instituições acadêmicas e científicas, revelando uma representação visual e quantitativa das interações dentro deste ecossistema. Dentro deste mesmo contexto, objetivando estabelecer um retrato da rede colaboração na perspectiva da produção técnica, nesta seção, é apresentada uma análise baseada em rede de colaboração.

A rede de colaboração será composta por vértices (nós), que representam individualmente cada inventor, e por arestas, que indicam as colaborações estabelecidas entre esses inventores no contexto do desenvolvimento de patentes. Para a construção da referida rede, adotou-se como critério de ligação a coautoria de patentes: dois inventores são considerados colaboradores quando constam conjuntamente em um mesmo pedido de patente. Por exemplo, considere a patente “X”, na qual estão listados os inventores “A” e “B”, e a patente “Y”, que apresenta como inventores “A” e “C” ([Figura 55](#)).

Figura 55 – Exemplo de rede de colaboração estabelecida entre três inventores

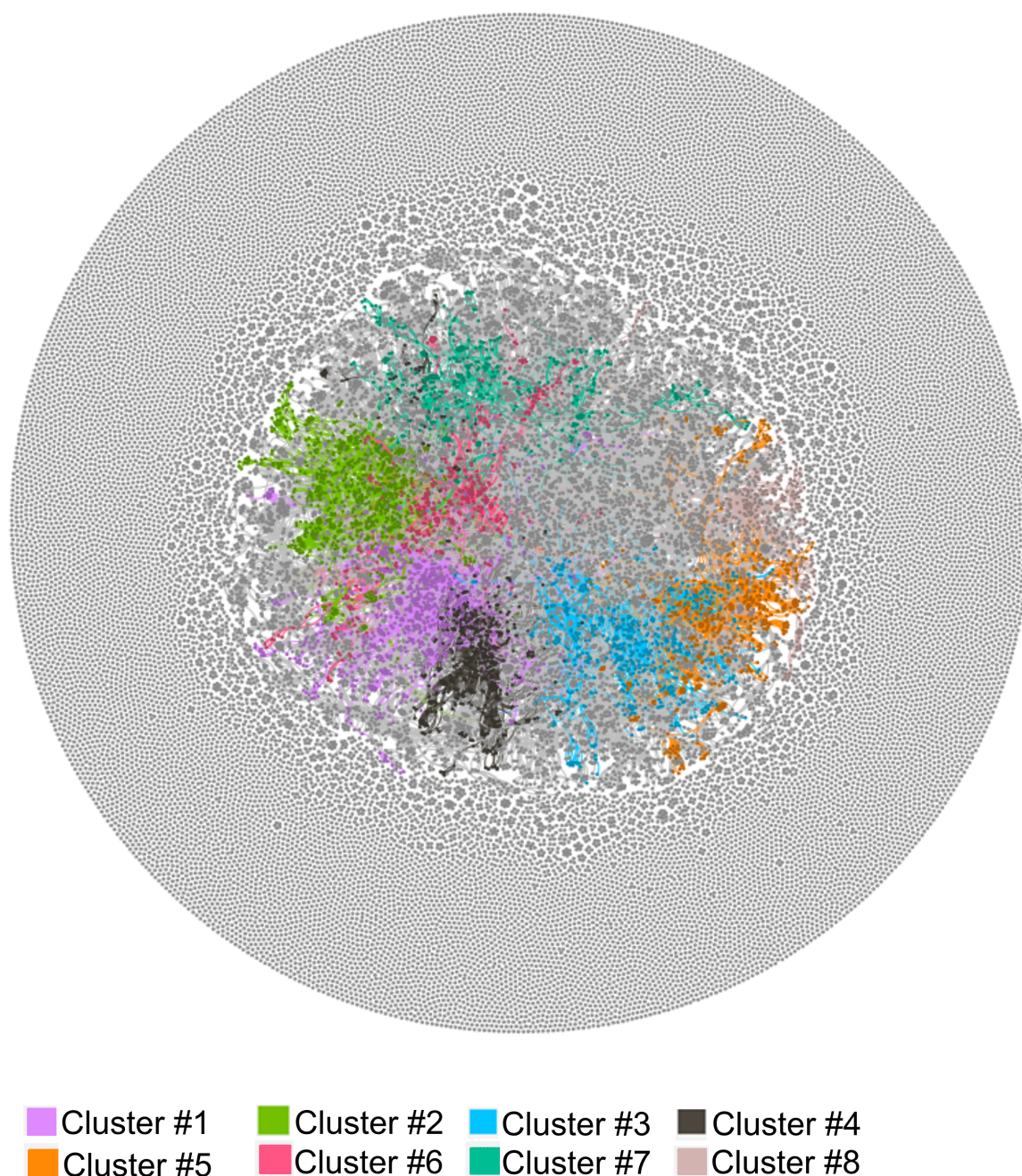


Fonte: Elaboração do autor

Conforme a figura, nesse cenário, o inventor “A” estabelece colaborações tanto com “B” quanto com “C”, resultando na criação de duas arestas: (A, B) e (A, C). Contudo, como os inventores “B” e “C” não compartilham entre si nenhuma patente em comum, não há formação de aresta entre eles, ou seja, não se estabelece uma relação direta de colaboração.

Dentro deste contexto, a [Figura 56](#) apresenta a visão geral da rede de colaboração entre os proponentes de patentes. Somente o conjunto patentes certificadas registradas na Plataforma Lattes foi considerado para a construção da rede.

Figura 56 – Visão geral da rede de colaboração



Fonte: Elaboração do autor

Para organizar visualmente os vértices e arestas da rede apresentada, foi utilizado o algoritmo Fruchterman-Reingold, uma técnica amplamente empregada para a disposição de grafos de maneira compreensível e visualmente agradável. Proposto por Thomas Fruchterman e Edward Reingold em 1991, esse algoritmo busca posicionar os vértices de forma que aqueles conectados por arestas curtas fiquem próximos uns dos outros, enquanto vértices não conectados ou conectados por arestas longas sejam posicionados mais distantes. Essa abordagem resulta em uma distribuição circular dos vértices, posicionando os elementos de

maior centralidade no centro da esfera (YUE et al., 2018).

Para o processamento computacional da rede, foi utilizado o *software* Gephi<sup>3</sup>, versão 0.10, que oferece uma interface amigável e recursos avançados para a análise e visualização de redes.

A rede apresentada pela a Figura 56, possui 35.072 vértices que representam cada inventor e 45.201 arestas, que estabelecem as colaborações entre os inventores para proposição das patentes. A detecção de comunidades na rede de colaboração entre inventores permitiu identificar oito agrupamentos principais, o que sugere a existência de subestruturas colaborativas coesas. Esses clusters representam conjuntos de inventores que mantêm vínculos mais intensos entre si do que com os demais participantes da rede, o que pode estar relacionado a fatores como especialização temática, filiação institucional, localização geográfica ou participação em redes formais de inovação. A presença dessas comunidades revela uma modularidade significativa na rede, evidenciando padrões de colaboração especializados e relativamente compartimentalizados.

Ao realizar a análise de uma rede de colaboração, é possível avaliar a produtividade e a influência de indivíduos e instituições por meio de métricas. As métricas de rede servem para quantificar e analisar várias propriedades e características de uma rede, permitindo que os pesquisadores e analistas compreendam melhor sua estrutura, comportamento e propriedades emergentes. A Tabela 16 apresenta as métricas da rede apresentada na Figura 56.

Tabela 16 – Estatísticas da rede de colaboração entre os pesquisadores

| Estatística                  | Valor     |
|------------------------------|-----------|
| Vértices                     | 35.072    |
| Arestas                      | 45.201    |
| Grau médio                   | 2,578     |
| Diâmetro da rede             | 28        |
| Componentes conectados       | 19.442    |
| Modularidade                 | 0,939     |
| Densidade                    | 0,0000734 |
| Comprimento médio de caminho | 8,568     |

A seguir, é discutida a análise estrutural da rede apresentada pela Tabela 16:

1. **Vértices = 35.072:** Reflete o número total de vértices na rede, neste caso, inventores individuais. Indica que a rede colaboração entre os inventores trata-se de uma rede relativamente grande, que envolve 35 mil inventores que, em algum momento, participaram de pedidos de patentes em conjunto com pelo menos um outro inventor;

<sup>3</sup> <<https://gephi.org/>>

2. **Arestas = 45.201**: Representa o total de colaborações diretas entre os inventores. Logo, existem 45.201 colaborações diretas estabelecidas a partir da coautoria de patentes;
3. **Grau médio = 2,578**: Indica que, em média, cada inventor colaborou diretamente com pelo menos duas outras pessoas. Este fato reflete um baixo grau de colaboração, o que indica uma rede mais fechada ou com parcerias pontuais, em vez de uma rede densa e interconectada;
4. **Diâmetro da rede = 28**: O diâmetro da rede indica a maior distância (em número de arestas) entre quaisquer dois vértices conectados na rede. Assim, a maior distância entre dois inventores conectados na rede é de 28 passos. Um valor relativamente alto, o que indica que a rede não é muito compacta. Existem caminhos longos entre certos grupos de inventores, além de vários componentes desconectados ou com apenas uma conexão;
5. **Componentes conectados = 19.442**: Apresenta o número de componentes conexos, ou seja, indica sub redes em que os vértices estão ligados, direta ou indiretamente. Esta métrica aponta que a rede possui muitos componentes separados, o que significa que grande parte dos inventores está em pequenas redes colaborativas, sendo um forte indicador de baixa coesão global da rede, indicando que algumas colaborações ocorrem em ambientes isolados;
6. **Modularidade = 0,939**: A modularidade mede a tendência da divisão da rede em comunidades ou clusters, variando entre 0 e 1. Logo, o valor 0,939 é considerado muito alto, o que indica que os inventores estão fortemente agrupados em comunidades com poucas conexões externas. Isso indica que existem grupos muito coesos internamente, mas com pouca interação entre grupos diferentes;
7. **Densidade = 0,0000734**: Esta métrica indica a proporção entre o número de arestas existentes e o número máximo possível de arestas. Uma densidade igual a 0,0000734 indica que a rede estabelecida entre os inventores é extremamente esparsa: pouquíssimas conexões em relação ao número total possível. Esta baixa densidade reforça a ideia de que a maioria dos inventores colabora apenas com um pequeno grupo;
8. **Comprimento médio do caminho = 8,568**: O comprimento médio do caminho indica a média das menores distâncias entre todos os pares de vértices conectados. Este indicador aponta que, em média, é necessário passar por cerca de 8,6 inventores para ir de um inventor a outro dentro da mesma componente conexa.

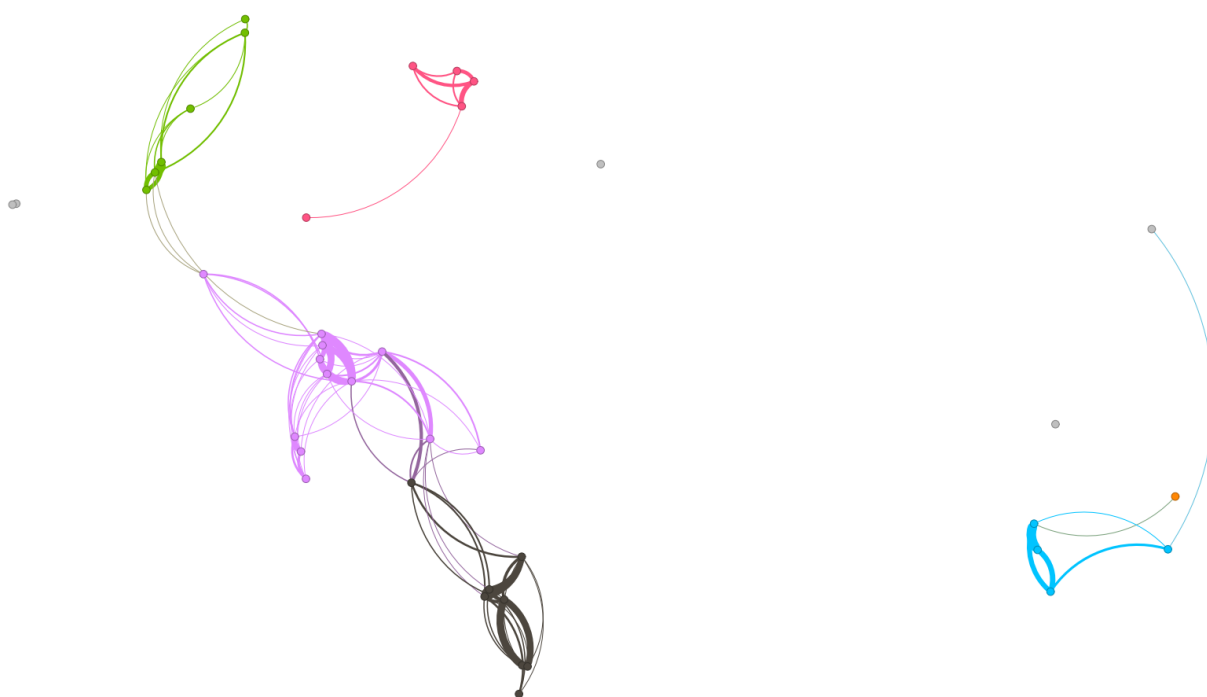
Em síntese, a rede de colaboração estabelecida entre inventores apresenta-se como uma rede representativa, refletida pelo elevado número de vértices e arestas, o que indica um volume significativo de dados e potenciais interações. No entanto, a análise das métricas revela um baixo nível de colaboração efetiva entre os inventores, uma vez que o grau médio da rede aponta que cada ator se conecta, em média, a apenas dois outros. Trata-se, portanto, de uma rede com pouca densidade e altamente fragmentada, caracterizada pela presença

de diversos clusters bem definidos, mas com reduzida interação entre eles, o que sugere a formação de comunidades com pouca colaboração. Essa configuração confere à rede um perfil extremamente esparsa, com baixa densidade de conexões e elevada modularidade, fatores que contribuem para uma difusão lenta de informações e conhecimentos ao longo de sua estrutura. Em contextos de inovação tecnológica, tal fragmentação pode limitar o potencial de disseminação de saberes e o aproveitamento da dinâmica entre diferentes grupos de pesquisa ou áreas técnicas.

O grau de um vértice representa a quantidade de arestas que estão conectadas a ele. Em outras palavras, quanto maior o grau (número de colaboradores) mais colaborativo é o indivíduo na rede. Na rede apresentada, o grau médio da rede é 2,578, ou seja, em média um inventor colabora com outros dois inventores na rede. Neste sentido, o inventor mais colaborativo é o “PESQUISADOR A” (foi atribuído um nome fictício para preservar sua privacidade) com 109 colaborações e 34 depósitos de patentes (certificadas), ele possui pós-doutorado, é bolsista de Produtividade em Pesquisa do CNPq - Nível 1A e sua idade científica é 19 anos.

Ampliando este cenário, ao selecionar os inventores com no mínimo 50 colabores (grau), obtém-se um total de 41 indivíduos. A [Figura 57](#) apresenta a sub-rede estabelecida entre eles.

Figura 57 – Sub-rede de entre os pesquisadores mais colaborativos



Fonte: Elaboração do autor

É possível observar que, em sua maioria, os inventores colaboram entre si, apesar de serem os mais colaborativos, os mesmo colaboram com menos de 1% dos indivíduos

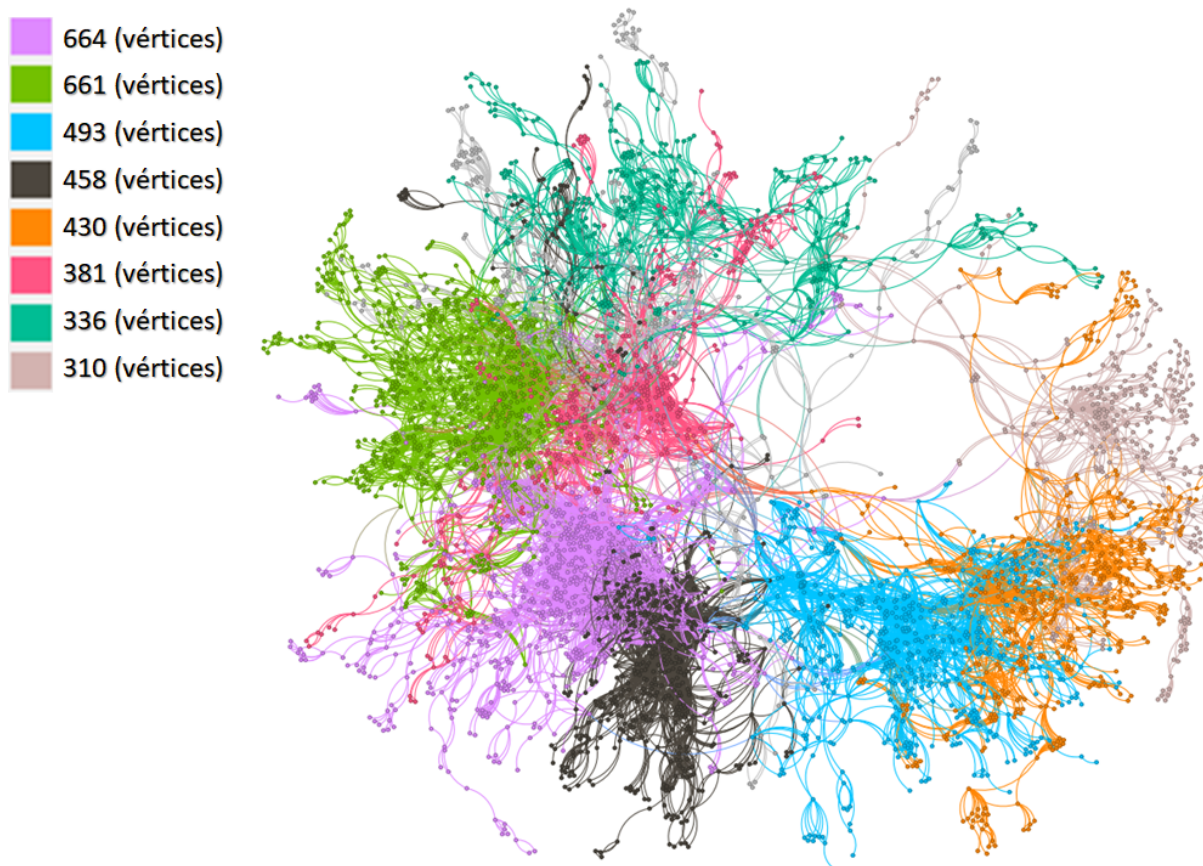
da rede. Entretanto, tais inventores representam um papel fundamental na construção e manutenção dessa rede. Eles desempenham um papel crucial no estabelecimento de conexões, na disseminação de informações e na promoção da cooperação entre os membros da rede. Em um contexto mais amplo, atuam como conectores, ligando indivíduos de diferentes partes da rede.

Em relação ao diâmetro, essa métrica define a distância mais longa entre quaisquer dois vértices da rede. A rede apresentada pela [Figura 56](#) tem o diâmetro igual a 28, ou seja, o caminho mais longo para relacionar dois indivíduos na rede é passando por 28 arestas. O diâmetro é uma métrica interessante para buscar entender a conectividade de uma rede de colaboração, pois quanto menor for o diâmetro da rede mais coesa é a rede, o que significa que os membros da rede estão mais próximos uns dos outros em termos de conexões. Isso pode facilitar a colaboração e a troca de conhecimento, uma vez que os indivíduos estão mais conectados.

Ainda explorando a conectividade da rede, a métrica componentes conectados, calculado pelo algoritmo *Robert Tarjan, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972)*, determina o número de grupos de arestas conectadas, ou seja, um conjunto de indivíduos no qual há pelo menos um caminho de conexão entre todos eles. Logo, a rede apresentada pela [Figura 56](#) possui 19.442 grupos, onde é possível estabelecer uma colaboração entre todos os indivíduos, sendo o maior grupo com 10.958 indivíduos (caracterizado como a componente gigante) e o menor grupo com apenas 1 indivíduo. Importante destacar que dentre as 19.442 grupos um total de 17.657 grupos são compostas por apenas um único indivíduo. Tal fator está relacionado ao fato que a classificação do algoritmo não conseguiu relacionar este indivíduo com nenhum outro dentro da rede.

Neste contexto, a métrica de modularidade verifica se uma rede pode ser dividida em comunidades modulares; quanto maior o valor da modularidade, ou seja, mais próximo de um, indica uma rede composta por uma complexa estrutura de comunidades internas. A rede apresentada possui modularidade igual a 0,939, calculado pelo algoritmo *Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000*. Neste sentido a, [Figura 58](#) apresenta uma sub-rede formada com as maiores comunidades identificadas.

Figura 58 – Sub-rede com as maiores comunidades identificadas



Fonte: Elaboração do autor

Na [Figura 58](#), são apresentadas somente as oito maiores comunidades presentes na rede, identificadas por cores distintas. Entretanto, a rede possui um total de 19.501 comunidades, a existência de muitas comunidades evidencia que os indivíduos tendem a selecionarem um determinado perfil de indivíduo para colaborar na proposição de patentes, por diversos motivos, tais como, similaridade de áreas de atuação, atuarem na mesma instituição, serem orientador e orientandos.

Ademais, a densidade da rede com valor menor que zero enfatiza a hipótese de uma rede composta por diversas comunidades, pois a densidade de uma rede é uma métrica que descreve o quão conectado são os vértices em relação ao número máximo de arestas possíveis. Em outras palavras, a densidade fornece uma indicação de quão perto a rede está de ser completamente conectada. E, por fim, a métrica de comprimento de caminho médio, calculado pelo algoritmo *Ulrik Brandes, A Faster Algorithm for Betweenness Centrality, in Journal of Mathematical Sociology 25(2):163-177, (2001)*, aponta que, na rede apresentada, em média, a distância entre dois indivíduos é de 8,5 colaborações.

Conforme o exposto, as métricas da rede aprontam uma rede complexa composta por várias comunidades internas. A fim de investigar este fenômeno, as arestas formadas

pelas colaborações podem conter informações relevantes. A cada aresta, é atribuído um peso que enfatiza a intensidade do relacionamento entre os vértices, ou seja, a cada patente depositada em colaboração entre os indivíduos aumenta o peso dessa colaboração. Neste sentido, 81,45% das colaborações foram realizadas uma única vez entre os indivíduos. Apenas 18,55% possuem grau superior a um, resultado de mais do que uma colaboração entre dois indivíduos, sendo que a colaboração que mais se repetiu, ocorreu 31 vezes. O “PESQUISADOR B” e “PESQUISADOR C” depositaram 31 patentes em conjunto. Analisando seus currículos na Plataforma Lattes, ambos são bolsistas de produtividade do CNPq.

Conforme apresentado nesta seção, a análise de redes de colaboração é crucial para entender e avaliar as interações entre os inventores. Além disso, se torna fundamental para promover a inovação e a criatividade, ao facilitar a identificação de oportunidades de colaboração que podem levar a novas soluções.

## 8 Considerações Finais

Este capítulo final consolida a trajetória de pesquisa desenvolvida ao longo desta tese. Após a apresentação detalhada do problema, da fundamentação teórica, da metodologia empregada, do processo de desenvolvimento e da análise dos resultados obtidos com a aplicação do protocolo de coleta e tratamento de dados bibliográficos de patentes, este momento é dedicado à síntese e à reflexão sobre o impacto do trabalho.

O capítulo está estruturado em três seções centrais: a [Seção 8.1](#) apresenta a conclusão formal, revisitando os objetivos e destacando as principais contribuições científicas e tecnológicas alcançadas; a [Seção 8.2](#) discute as perspectivas para trabalhos futuros, apontando novos caminhos para a evolução desta pesquisa; por fim, a [Seção 8.3](#) elenca as publicações acadêmicas resultantes deste esforço investigativo, que atestam a disseminação e a validação dos achados pela comunidade científica.

### 8.1 Conclusão e principais contribuições

A inovação e a competitividade dos países e empresas em um cenário globalizado, voltado ao compartilhamento de informações e conhecimento, estão diretamente associadas aos direitos da Propriedade Intelectual. Empresas e Universidades sentem uma crescente necessidade em assegurar uma proteção jurídica para produtos e/ou tecnologias que desenvolvem, contribuindo para o contínuo e crescente montante de depósito de patentes, constituindo de tal forma, um imponente conjunto de dados sobre o desenvolvimento tecnológico. Oportunamente, mediante a análise de documentos de patentes, pode-se investigar o estado da arte, tendências tecnológicas, identificação de concorrentes, avaliação de liberdade de operação, definir estratégias de investimentos e inovação, entre outras perspectivas sobre a propriedade intelectual.

Diversas organizações têm usado a análise de dados de patentes como um diferencial competitivo para se manter no mercado, identificando seus principais concorrentes, bem como os atores responsáveis de áreas de interesse da organização, fator importante para estabelecer colaborações e parcerias. Ainda, para além das organizações, as Instituições de Ensino e Pesquisa têm assumido, ao longo dos anos, um papel cada vez mais relevante no desenvolvimento tecnológico, agregando o empreendedorismo aos seus princípios de ensino, pesquisa e extensão, convertendo parte do conhecimento científico desenvolvido nos laboratórios em publicações de patentes.

Durante o curso da presente investigação, foi evidenciado que as patentes se configuram como fontes de informação privilegiadas, oferecendo dados precisos e, em

muitos casos, singulares. Entretanto, sua manipulação é frequentemente obstaculizada por diversos fatores. Entre estes, destacam-se: a vasta quantidade de patentes disponíveis e suas atualizações contínuas; a adoção de uma linguagem técnica e jurídica específica na redação dos documentos patentários; a dificuldade em localizar as patentes pertinentes para uma análise específica, especialmente devido à necessidade de empregar termos-chave precisos e estratégias de busca avançadas; as restrições para a recuperação de amplos conjuntos de dados em repositórios de patentes de acesso aberto.

Este estudo se propôs a contribuir de maneira significativa com o avanço e a disseminação de pesquisas em Propriedade Intelectual, com foco na coleta e gestão de dados bibliográficos de patentes. Para isso, foi desenvolvido e validado um protocolo sistemático de coleta e tratamento de grandes volumes de dados bibliográficos de patentes, buscando superar as limitações existentes em repositórios de acesso aberto. A pesquisa utilizou uma abordagem multidisciplinar, combinando elementos da Ciência da Computação e da Ciência da Informação, e culminou na criação de um protocolo estruturado que integra técnicas de organização e recuperação da informação, inteligência artificial, bibliometria e patentometria.

Assim, foram abordadas as problemáticas pertinentes ao processo de gestão de grandes conjuntos de dados bibliográficos de patentes, através de levantamento documental e bibliográfico de informações disponíveis tanto em formato digital quanto impresso, como livros, artigos em periódicos e repositórios de patentes.

Os principais resultados desta pesquisa são refletidos na construção de um protocolo de coleta e tratamento de dados bibliográficos de patentes, o Protocolo CTDP. Este protocolo foi projetado para orientar a coleta, o tratamento e o processamento de dados bibliográficos de patentes de diversas magnitudes, abrangendo desde conjuntos menores e específicos até volumes massivos de informações. Sua concepção representa uma resposta direta e fundamental à questão de pesquisa deste estudo e em conformidade com a hipótese estabelecida.

A definição de onze critérios para a seleção de repositórios revelou-se um passo fundamental para formalizar e orientar o pesquisador no processo de escolha do repositório de patentes mais apropriado para a recuperação de dados bibliográficos. Em um cenário onde a consistência e a completude dos dados são imperativos para a validade dos resultados de pesquisa, a decisão sobre a fonte de dados primária assume um papel crucial.

A relevância desses critérios reside na mitigação de riscos inerentes à coleta de dados em ambientes abertos. Uma escolha inadequada de um repositório pode inviabilizar a obtenção de um conjunto de dados representativo para análise e comprometer a validade e a confiabilidade dos resultados da pesquisa. Isso se deve ao fato de que diferentes repositórios de patentes, como o Espacenet, Patentscope ou o Google Patents, embora ofereçam acesso público, podem apresentar variações substanciais em termos de cobertura,

disponibilização de dados, formatos de exportação, entre outros. Tais variações, se não consideradas por um conjunto claro de princípios norteadores, podem comprometer a pesquisa.

Portanto, os critérios estabelecidos atuam como um balizador, assegurando que a seleção do repositório não seja uma decisão arbitrária, mas sim um processo fundamentado em critérios objetivos e transparentes. Esta abordagem sistemática contribui para garantir a reprodutibilidade da pesquisa.

De igual modo, as estratégias de coleta estabelecidas neste estudo delinearam um protocolo que possibilita a recuperação de conjuntos amplos de dados bibliográficos sem incorrer no descumprimento dos critérios e termos de serviço estabelecidos pelos repositórios de patentes de acesso aberto. Este é um ponto importante, dado que a adesão às políticas de uso dos provedores de dados, como a limitação de requisições ou a proibição de *scraping* agressivo, é obrigatório para adquirir o conjunto de dados.

Assim sendo, a relevância dessa conformidade com os repositórios está na capacidade de acesso aos dados a longo prazo. Estratégias que desconsideram os termos estabelecidos pelos repositórios podem resultar em bloqueios temporários de acesso ou, em casos mais graves, na revogação das permissões de uso. Portanto, a concepção de um sistema que contempla essas restrições, ao invés de ignorá-las, ressalta a metodologia eticamente responsável empregada neste trabalho.

Não obstante a complexidade da coleta, as abordagens metodológicas definidas para o armazenamento, a transformação e a deduplicação dos dados bibliográficos de patentes culminaram na criação de uma base de dados local. Esta base de dados viabiliza análises de diversas perspectivas, desde estudos bibliométricos básicos até investigações mais complexas envolvendo o processamento de linguagem natural e a identificação de tendências tecnológicas.

Adicionalmente, a interoperabilidade dessa base de dados local com o catálogo de fontes complementares estabelecido neste estudo, com dados do Instituto Nacional da Propriedade Industrial (INPI) e registros da Plataforma Lattes (CNPq), representa um avanço significativo. Essa integração enriquece o conjunto de dados bibliográficos das patentes, com informações contextuais sobre os pesquisadores, suas redes de colaboração e a relevância de suas patentes. A capacidade de correlacionar dados de patentes com informações curriculares de pesquisadores, por exemplo, abre novas perspectivas para a investigação da relação entre a produção tecnológica e a produção acadêmica.

É importante ressaltar que a base de dados gerada pode ser facilmente integrada a diversas ferramentas de análise e mineração de dados, tais como Gephi, Carrot e Microsoft Power BI. Essa integração é viabilizada pelo fato da base de dados ser estruturada sob o modelo relacional, o que possibilita a conexão com qualquer ferramenta que suporte a

interação com bancos de dados relacional.

Outra contribuição relevante deste estudo consiste na implementação de estratégias de pesquisa semântica de patentes, bem como na classificação e agrupamento de documentos de patentes utilizando *Large Language Models* (LLMs). Essa abordagem inovadora visa simplificar a linguagem complexa e frequentemente técnica que caracteriza a redação desses documentos.

Comumente, linguagem técnica e jurídica presente nos documentos de patentes contribui para a subutilização de documentos de patentes. Ao empregar LLMs, este estudo oferece uma solução para simplificar essa linguagem, transformando dados textuais complexos em informações mais acessíveis. Isso otimiza o processo de recuperação e também viabiliza o agrupamento de patentes, revelando padrões e tendências. Consequentemente, essa inovação tem o potencial de acelerar processos de prospecção tecnológica, beneficiando tanto pesquisadores quanto tomadores de decisão em ambientes de inovação.

A eficácia e aplicabilidade do Protocolo CTDP foram empiricamente validadas por sua aplicação na coleta de patentes brasileiras disponíveis no repositório Espacenet, depositadas no período de 1900 a 2024. Essa abordagem demonstrou a capacidade do Protocolo CTDP de lidar com a complexidade inerente à coleta de grande conjuntos de dados de patentes. A validação em um contexto real comprovou sua utilidade como ferramenta para investigações em inovação e propriedade intelectual.

As implicações práticas deste estudo se materializaram na implementação do conjunto ferramental Pytente, uma ferramenta que implementa todas as diretrizes estabelecidas pelo o Protocolo CTDP. Sendo uma ferramenta de código aberto (*Open Source*), oferece uma solução acessível e flexível. Sua compatibilidade com os principais sistemas operacionais, incluindo Windows, Linux e macOS, expande consideravelmente sua aplicabilidade e potencial de adoção na comunidade científica e tecnológica. O caráter *open source* do Pytente fomenta a colaboração e o aprimoramento contínuo da ferramenta e promove a transparência e a reprodutibilidade metodológica, aspectos fundamentais em pesquisas acadêmicas.

Apesar das relevantes contribuições, é importante reconhecer que a validação do protocolo se concentrou em um único repositório (Espacenet) para patentes brasileiras, o que pode limitar a generalização dos resultados para outros contextos e tipos de documentos de propriedade intelectual. Contudo, essa limitação não invalida a metodologia proposta, mas aponta para a necessidade de pesquisas futuras.

Em síntese, esta Tese oferece uma contribuição substancial para o campo da Propriedade Intelectual e da Ciência da Informação, ao prover um protocolo sistemático e validado para a coleta e o tratamento de dados bibliográficos de patentes. Espera-se que este trabalho facilite e inspire futuras pesquisas baseadas em dados de patentes e contribua

para o avanço do conhecimento em inovação e desenvolvimento tecnológico.

## 8.2 Trabalhos futuros

Este trabalho contribuiu significativamente para o entendimento e a análise de grandes volumes de dados de patentes, ao propor e validar uma metodologia para a construção de uma base de dados local, integrando informações de repositórios de acesso aberto e fontes complementares. A pesquisa demonstrou a viabilidade e a eficácia do protocolo desenvolvido. No entanto, considerando a complexidade e a amplitude do tema, abre-se um vasto horizonte de oportunidades para expandir e aprofundar os resultados aqui apresentados.

Nesta seção, são discutidas possíveis extensões e aprimoramentos que podem ser explorados em pesquisas futuras. Tais investigações visam superar as limitações identificadas no estudo atual e ampliar o impacto de suas contribuições para a comunidade científica e tecnológica.

A primeira oportunidade de estudo futuro consiste na “Extensão e Validação do Protocolo em Múltiplos Repositórios”. A presente Tese estabeleceu e validou o protocolo de coleta utilizando o Espacenet como fonte primária de dados. Uma extensão de grande relevância consiste na implementação e adaptação do processo para outros repositórios de patentes de acesso aberto, como The Lens e Google Patents. Cada repositório possui uma arquitetura de dados, uma API e uma cobertura geográfica/temporal distinta. Adaptar o protocolo para múltiplas fontes permitiria testar a generalidade e adaptabilidade. Este passo é importante para validar a metodologia como um framework metodológico flexível. Além disso, a integração de dados de diferentes repositórios poderá mitigar vieses de cobertura e enriquecer a base de dados final, permitindo análises comparativas sobre a consistência da informação entre as principais plataformas globais.

A segunda oportunidade surge no “Aprimoramento da Extração de Informações com Aprendizado de Máquina”. Este estudo focou no tratamento de dados bibliográficos estruturados. Uma fronteira clara de avanço reside na exploração de técnicas de Aprendizado de Máquina (*Machine Learning*) para a extração e o tratamento semântico de informações contidas nos campos textuais de patentes (resumo, reivindicações e descrição). A aplicação de modelos de Processamento de Linguagem Natural (PLN), como a Reconhecimento de Entidades Nomeadas (NER), pode automatizar a identificação de elementos críticos como compostos químicos, tecnologias específicas, métricas de performance e nomes de organizações não normalizados. Modelos de Tópicos (*Topic Modeling*) podem ser empregados para classificar automaticamente patentes em categorias tecnológicas emergentes, como alternativa aos esquemas de classificação tradicionais. Tal avanço transforma a base de dados de um repositório bibliográfico para uma base de conhecimento rica e semanticamente

estruturada, habilitando análises de inteligência tecnológica de maior profundidade.

Na sequência, a “Certificação de Patentes na Plataforma Lattes com Inteligência Artificial” é outra oportunidade relevante de trabalho futuro. A integração com a Plataforma Lattes foi um diferencial deste trabalho. Contudo, o processo de associação entre um pesquisador e uma patente declarada em seu currículo ainda apresenta desafios de validação, dado que a informação é autodeclarada. Uma pesquisa futura de alto impacto é o desenvolvimento de um modelo de Inteligência Artificial para otimizar este processo de certificação. Um modelo baseado em redes neurais ou outras técnicas de aprendizado de máquina pode analisar o conteúdo semântico da patente e compará-lo com a produção científica (artigos, teses) e as áreas de atuação do pesquisador no Lattes. O resultado é um mapa inédito e confiável da contribuição de pesquisadores brasileiros para a inovação tecnológica, uma ferramenta de grande valor para agências de fomento, universidades e gestores de política científica e tecnológica.

Atualmente, a ferramenta resultante desta tese, denominada Pytente, opera por meio misto, onde, na interface Web, são elaboradas as estratégias e executados localmente, por meio de execução de scripts. Apesar de todo o tratamento realizado neste estudo para tornar o processo simples e acessível a todo tipo de usuário, o desenvolvimento de uma interface gráfica para a execução local torna a ferramenta ainda mais acessível a um público não especializado. Reafirmando, dessa forma, o compromisso com democratização do acesso à análise de dados de patentes.

E, por fim, a migração para uma plataforma 100% online (*Software as a Service - SaaS*) elimina a necessidade de instalação e processamento local, superando barreiras de hardware e configuração. Os usuários poderiam acessar a ferramenta de qualquer dispositivo, e o processamento de grandes volumes de dados seria centralizado em servidores de alto desempenho.

Em conjunto, as direções de pesquisa aqui apresentadas apresentam um vasto potencial de expansão deste trabalho. Cada sugestão utiliza o protocolo e a base de dados desenvolvidos como elemento fundamental. Fica evidente que as possibilidades desta tese não se encerram em suas conclusões; pelo contrário, ela estabelece as fundações para um programa de pesquisa de longo prazo, com potencial para transformar a maneira como a recuperação e análise de patentes é conduzida no ambiente acadêmico e tecnológico.

### 8.3 Publicações

A seguir, são apresentadas as publicações científicas desenvolvidas com base nos estudos realizados neste trabalho. Foram produzidos e publicados os seguintes artigos:

**Artigos publicados em anais de congressos:**

### 1. Collection and Processing of Data on Brazilian Technical Production

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; MASCARENHAS, Higor A. D.. Collection and Processing of Data on Brazilian Technical Production in Engineering Areas. *In: The XLIII Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE 2022)*. 43., 2022. **Anais....** Foz do Iguaçu, 2022.

### 2. Um retrato da produção técnica brasileira: análise baseada em registros de patentes na Plataforma Lattes

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Um retrato da produção técnica brasileira: análise baseada em registros de patentes na Plataforma Lattes. *In: Encontro Brasileiro de Bibliometria e Cientometria (8ª EBBC)*. 8., 2022. **Anais....** Maceió, 2022.

### 3. Patentes como fonte de dados para análise sobre a produção técnica

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues. Patentes como fonte de dados para análise sobre a produção técnica. *In: Anais do VI Workshop de Informação Dados e Tecnologia - 5ª WIDaT*. 5., 2022 **Anais....** Vitória: Universidade Federal do Espírito Santo – UFES, 2022. ISBN: 978-65-00-61990-4.

### 4. Estratégia para certificação de patentes em repositórios internacionais

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Estratégia para certificação de patentes em repositórios internacionais. *In: Anais do VI Workshop de Informação Dados e Tecnologia - 6ª WIDaT*. 6., 2023 **Anais....** DOI: <<https://doi.org/10.22477/vi.widat.34>>

### 5. Proposição de um framework de coleta e tratamento de dados sobre a produção técnica brasileira

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Proposição de um framework de coleta e tratamento de dados sobre a produção técnica brasileira. *In: XXIII Encontro Nacional de Pós-Graduação em Ciência da Informação - ENANCIB*. 23., 2023. **Anais....** São Cristóvão, 2023.

## 6. Protocolo de coleta de dados bibliográficos de patentes

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Protocolo de coleta de dados bibliográficos de patentes. *In: 9<sup>o</sup> Encontro Brasileiro de Bibliometria e Cientometria - 9<sup>o</sup> EBBC. 2024. Anais.... Brasília-DF, 2024.*

## 7. Coleta e integração de fontes de dados heterogêneas sobre patentes

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Coleta e integração de fontes de dados heterogêneas sobre patentes. *In: XXIV Encontro Nacional de Pós-Graduação em Ciência da Informação - ENANCIB. 24., 2024. Anais.... Vitória-ES, 2024.*

## 8. Framework de automação de coleta de dados bibliográficos de patentes em repositórios de acesso aberto

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Framework de automação de coleta de dados bibliográficos de patentes em repositórios de acesso aberto. *In: XIV Conferência Internacional BIREDIAL-ISTEC 2025., 2025. Anais.... Brasília-DF.*

## 9. Implementação de uma estratégia de busca semântica em conjuntos de dados bibliográficos de patentes

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Implementação de uma estratégia de busca semântica em conjuntos de dados bibliográficos de patentes. *In: XIV Conferência Internacional BIREDIAL-ISTEC 2025., 2025. Anais.... Brasília-DF.*

## 10. Temporal analysis of patents in the engineering field

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; MASCARENHAS, Higor A. D.. Collection and Processing of Data on Brazilian Technical Production in Engineering Areas. *In: The XLVI Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE 2025)., 2025. Anais.... Vitória-ES, 2025.*

## 11. Caracterização da produção técnica gerada por instituições de ensino superior: Um estudo de patentes

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Caracterização da produção técnica gerada por instituições de ensino superior: Um estudo de patentes. *In*: XXV Encontro Nacional de Pós-Graduação em Ciência da Informação - ENANCIB., 2025. **Anais....** Rio de Janeiro-RJ, 2025.

**12. Pytente - Framework de automação de coleta e tratamento de dados bibliográficos de patentes em repositórios de acesso aberto.**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Pytente - Framework de automação de coleta e tratamento de dados bibliográficos de patentes em repositórios de acesso aberto. *In*: 16ª Conferência lusófona Ciência Aberta - ConfOA 2025., 2025. **Anais....** Goiânia-GO.

**Artigos publicados em revistas:**

**1. Caracterização e análise da produção técnica no Brasil: um estudo baseado em registros de patentes**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Caracterização e análise da produção técnica no Brasil: um estudo baseado em registros de patentes. *In*: **Informação & Informação**. Londrina, v. 27, n. 3, p. 380–401, jul./set. 2022, DOI: 10.5433/1981-8920.2022v7n3p380.

**2. Collection and Integration of Patent Data for Analysis and Validation of Brazilian Technical Production**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Collection and Integration of Patent Data for Analysis and Validation of Brazilian Technical Production. *In*: **Mobile Networks and Applications**. 2023. <<https://doi.org/10.1007/s11036-023-02234-3>>

**3. Analisando a produção técnica brasileira: uma abordagem considerando registros de patentes**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues. Analisando a produção técnica brasileira: uma abordagem considerando registros de patentes.

*In: Revista Ibero-Americana De Ciência Da Informação*,. 2023. 16(1), 245–262.  
<<https://doi.org/10.26512/rici.v16.n1.2023.47597>>

### **Artigos publicados em capítulos de livro:**

#### **1. Collection and Integration of Patent Data for Analysis and Validation of Brazilian Technical Production**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de Carvalho Segundo. Collection and Integration of Patent Data for Analysis and Validation of Brazilian Technical Production. *In: DATA and Information in Online Environments*: Third EAI International Conference, DIONE 2022, Virtual Event, July 28-29, 2022, Proceedings. 1. ed. [S. l.]: Springer Cham, 2022. p. 39-48. ISBN 978-3-031-22323-5.

### **Colaborações/contribuições científicas**

#### **1. (2023-2025) Projeto de pesquisa "BrCris: Brazilian Current Research Information System"**

O *Brazilian Current Research Information System* (BrCris) é uma plataforma desenvolvida ao longo de aproximadamente dez anos em colaboração com instituições como CEFET-MG, Fiocruz, UFABC e PUC-RJ. Sua principal função é organizar e disponibilizar informações estratégicas sobre pesquisadores, projetos e instituições, promovendo o acesso aberto e alinhando-se aos princípios da Ciência Aberta. O sistema integra dados de múltiplas fontes, incluindo a Plataforma Lattes, a Capes e repositórios como DOAJ e Espacenet. Utilizando o Elasticsearch para buscas e a construção de painéis de indicadores, o BrCris possibilita uma visão abrangente do ecossistema nacional de pesquisa, apoiando processos de tomada de decisão e ampliando a compreensão sobre o cenário científico e tecnológico brasileiro.

#### **2. (2025-2027) Projeto de pesquisa "ServiçosPI: inovação e eficiência na recuperação de informações de propriedade intelectual do INPI"**

No Brasil, o Instituto Nacional da Propriedade Industrial (INPI) exerce um papel estratégico na proteção de ativos intangíveis, como patentes e marcas, fundamentais para o fortalecimento da economia baseada no conhecimento. Contudo, o atual sistema de busca do INPI, conhecido como Busca Web, encontra-se tecnicamente defasado, pois foi implementado há mais de dez anos utilizando tecnologias como Servlets/JSP, hoje consideradas obsoletas. Essas limitações impactam negativamente

o desempenho e a escalabilidade do sistema, sobretudo em momentos de alta demanda, como durante as publicações da Revista da Propriedade Industrial (RPI). Essa situação foi recentemente destacada pelo Supremo Tribunal Federal (STF) no julgamento da Ação Direta de Inconstitucionalidade (ADI) nº 5.529, que apontou a insuficiência tecnológica do INPI como um entrave ao exercício pleno de suas funções institucionais, prejudicando sua atuação no fomento à inovação e ao desenvolvimento tecnológico nacional (Supremo Tribunal Federal, 2020).

Em resposta a esses desafios, foi proposta uma cooperação entre o INPI e o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) para a execução do projeto de pesquisa "ServiçosPI: inovação e eficiência na recuperação de informações de propriedade intelectual". O objetivo central do projeto é o desenvolvimento de uma solução de busca moderna, escalável e eficiente, que atenda aos diversos perfis de usuários, tais como inventores, procuradores, pesquisadores, órgãos de controle e áreas internas do próprio INPI. A proposta contempla a implementação de funcionalidades como a disponibilização online da RPI, gestão de publicações de pedidos e registros, criação de áreas personalizadas para usuários, mecanismos de busca otimizados e integração com fontes externas, assegurando maior interoperabilidade, agilidade e precisão no tratamento e na recuperação de informações de propriedade intelectual.

### **Produção intelectual: software registrado**

#### **1. Pytente**

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; SEGUNDO, Washington Luís Ribeiro de Carvalho. Pytente. [S. l.], 20 maio 2025. BR512025002005-8. Disponível em: <https://pytentui.web.app>.

## Referências

- ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, Elsevier, v. 37, p. 3–13, 2014. Disponível em: <<https://doi.org/10.1016/j.wpi.2013.12.006>>. Acesso em: 24 dez. 2023. Citado 6 vezes nas páginas 32, 33, 46, 52, 64 e 65.
- ABRANTES, P. C. de et al. Estudo comparativo de bases gratuitas de patentes: patentscope (wipo), espacenet(epo), buscaweb (inpi/br). **Revista Tecnologia e Sociedade**, v. 18, n. 54, p. 125–142, 12 2022. Disponível em: <<https://revistas.utfpr.edu.br/rts/article/viewFile/14516/9108>>. Citado na página 27.
- ALMEIDA, N. G. N. de. A importância da metodologia científica através do projeto de pesquisa para a construção da monografia. **Folha de rosto em Biblioteconomia e Ciência da Informação**, v. 2, n. 1, p. 57–66, 2016. Citado na página 47.
- 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. Citado na página 1.
- ARISTODEMOU, F. L.; TIETZE. The state-of-the-art on intellectual property analytics (ipa): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (ip) data. **World Patent Information**, v. 55, p. 37–51, 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0172219018300103>>. Acesso em: 07 mai. de 2021. Citado na página 37.
- BARBOSA, D. B. **Uma Introdução À Propriedade Intelectual**. segunda. [S.l.]: Editora Lumen Juris, 2010. Disponível em: <<http://s-inova.ucdb.br/wp-content/uploads/biblioteca/Uma%20Introducao%20a%20Propriedade%20Intelectual.pdf>>. Acesso em: 13 abr. de 2021. Citado na página 8.
- 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, set. 2016. Disponível em: <<https://lume.ufrgs.br/handle/10183/179853>>. Acesso em: 24 mar. de 2021. Citado 3 vezes nas páginas 23, 24 e 26.
- BRASIL. Código de propriedade industrial - lei 9279/96. **Diário Oficial da União**, mai. 1996. Citado 2 vezes nas páginas 12 e 13.
- BRASIL. **LEI Nº 9.279, DE 14 DE MAIO DE 1996**. BRASIL, 1996. Disponível em: <[http://www.planalto.gov.br/ccivil\\_03/leis/l9279.htm](http://www.planalto.gov.br/ccivil_03/leis/l9279.htm)>. Acesso em: 23 de nov. de 2023. Citado 2 vezes nas páginas 22 e 23.
- BRASIL. **Bases de patentes online**. 2015. Disponível em: <<https://www.gov.br/inpi/pt-br/assuntos/informacao/bases-de-patentes-online>>. Citado na página 58.
- BRASIL. **Legislação geral e pareceres**. Brasil, 2020. Disponível em: <<https://www.gov.br/inpi/pt-br/backup/centrais-de-conteudo/legislacao>>. Acesso em: 16 de nov. de 2023. Citado na página 10.

BRASIL. **Patentes - Instituto Nacional da Propriedade Industrial**. Brasil, 2020. Disponível em: <<https://www.gov.br/inpi/pt-br/servicos/perguntas-frequentes/patentes>>. Acesso em: 20 de nov. de 2023. Citado 2 vezes nas páginas 11 e 12.

BRASIL. **Guia Básico**. Brasil, 2021. Disponível em: <<https://www.gov.br/inpi/pt-br/servicos/patentes/guia-basico>>. Acesso em: 20 de nov. de 2023. Citado na página 13.

BRASIL. **Oasisbr passa a ser utilizado para certificação de Teses e Dissertações no Currículo Lattes: Teses e dissertações do Lattes agora contam com o selo de certificação de autenticidade do Oasisbr**. 2022. Disponível em: <<https://www.gov.br/cnpq/pt-br/assuntos/noticias/cnpq-em-acao/plataforma-curriculo-lattes-passa-a-ser-certificada-pelo-oasisbr>>. Acesso em: 04 mai. de 2023. Citado na página 98.

BRASIL. **Classificação de patentes**. Brasil, 2023. Disponível em: <<https://www.gov.br/inpi/pt-br/servicos/patentes/classificacao-de-patentes>>. Acesso em: 03 de mar. de 2020. Citado na página 15.

BRASIL. **Pedidos de patentes de residentes crescem 10% em 2023**. Brasil, 2024. Disponível em: <<https://www.gov.br/inpi/pt-br/central-de-conteudo/noticias/pedidos-de-patentes-de-residentes-crescem-10-em-2023>>. Acesso em: 16 de dez. de 2024. Citado na página 118.

BRITO, L. A. G. C. de; QUONIAM, J. P.; MENA-CHALCO. Exploração da plataforma lattes por assunto: proposta de metodologia. **Transinformação**, v. 28, n. 1, p. 77–86, abr. 2016. FapUNIFESP (SciELO) <<http://dx.doi.org/10.1590/2318-08892016002800006>>. Citado na página 144.

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://doi.org/10.1590/S1413-99362011000400003>>. Acesso em: 08 mai. de 2021. Citado na página 34.

CARVALHO, M. R. d. M. C. B. C. C. B.; SANTOS. A classificação internacional de patentes: Descrição e importância. **GEINTEC**, v. 09, n. 01, p. 4798–4808, jan. 2019. Citado na página 18.

CATIVELLI, A. S. **Patentes das Universidades Públicas Brasileiras: Análise das Concessões**. Dissertação (Dissertação do Mestrado Profissional em Gestão de Unidades de Informação) — Universidade do Estado de Santa Catarina, 2016. Citado na página 41.

CNA. **PIB do agronegócio avança no trimestre e acumula alta de 9,81% no primeiro semestre de 2021**. [S.l.], 2021. Disponível em <[https://www.cnabrazil.org.br/assets/arquivos/boletins/dtec.pib\\_jun\\_2021.13set2021vff.pdf](https://www.cnabrazil.org.br/assets/arquivos/boletins/dtec.pib_jun_2021.13set2021vff.pdf)> Acesso em: 05 de jan. de 2022. Citado na página 121.

CUNHA, K. C. T.; VOLPATO, G.; PEDRON, C. D. Desenvolvimento de um protocolo de revisão sistemática de bases patentárias: A complexidade da linguagem de acesso. In: **Anais do Workshop de Informação, Dados e Tecnologia - WIDaT**. Brasília: [s.n.], 2023. v. 6. Disponível em: <<https://labcotec.ibict.br/widat/index.php/widat2023/article/view/61>>. Acesso em: 12 jan. 2024. Citado na página 44.

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. Citado 4 vezes nas páginas 34, 35, 36 e 37.

DANILO ZAMPRONIO. **Como ler uma patente**. O Consultor em Patentes, 2021. Disponível em: <<https://oconsultorempatentes.com/como-ler-uma-patente/>>. Acesso em: 23 de nov. de 2023. Citado na página 21.

DATE, C. J. **Introdução a Sistemas de Banco de Dados**. eighth. [S.l.]: Elsevier, 2003. Citado na página 77.

DIAS, T. H. J. T. M. R.; MOREIRA, P. M.; DIAS. Caracterização e análise das redes de colaboração científica dos bolsistas de produtividade em pesquisa do cnpq. **BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING**, n. 7, p. 64–75, 2018. Natal. Anais. Natal: Sbc. Citado na página 144.

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. Citado 5 vezes nas páginas 77, 93, 94, 145 e 148.

DIGIAMPIETRI, L. A. **Análise da Rede Social Acadêmica Brasileira**. Tese (LIVRE DOCÊNCIA - Informação e Tecnologia) — Escola de Artes Ciências e Humanidades da Universidade de São Paulo, 2015. São Paulo. Citado na página 144.

DIGIAMPIETRI, L. A. et al. Minerando e caracterizando dados de currículos lattes. **Anais do brazilian workshop on social network analysis and mining (BRASNAM)**, 2012. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/6868>>. Acesso em: 05 jun. de 2023. Citado na página 98.

ESPECENET. **Espacenet patent search**. 2023. Disponível em: <<https://worldwide.espacenet.com/patent/>>. Acesso em: 07 de out. de 2023. Citado na página 26.

FAGUNDES, P. A. d. A. M. C. et al. Perfil tecnológico da csu: Um estudo patentométrico. **Revista de Administração e Inovação (RAI)**, v. 11, n. 1, p. 276–294, mar. 2014. Citado na página 37.

FERREIRA, A. B. de H. **Dicionário Aurélio**. [S.l.]: Editora Positivo, 2010. Citado na página 8.

FRANÇA, R. O. Patente como fonte de informação tecnológica. **Perspectivas em Ciência da Informação**, v. 2, n. 2, 2007. Disponível em: <<https://periodicos.ufmg.br/index.php/pci/article/view/23241>>. Citado na página 40.

GIL, A. C. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas, 1991. Citado na página 47.

GIUSTI, L. J. M. et al. Sustentabilidade na engenharia de produção: um estudo bibliométrico de 2001 a 2011. In: **Simpósio de Engenharia de Produção**. São Paulo: [s.n.], 2011. v. 18. Citado na página 35.

GOOGLE. **About Google Patents Overview**. Google Patents, 2023. Disponível em: <[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)>. Acesso em: 19 de dez. de 2023. Citado na página 28.

HQIT, H. Q. I. T. C. L. **A patent database management analysis method**. 2017. Depositante: Hefei Qingzhuo Information Technology Co. Ltd. CN106407394A. Depósito: 19 set. 2016. Publicação: 15 fev. 2017. Citado na página 2.

IBM, I. B. M. C. **Processing Large Machine Learning Datasets**. 2022. Depositante: International Business Machines Corporation. US2022075761A1 . Depósito: 8 nov. 2020. Publicação: 10 fev. 2022. Citado na página 2.

INPI. **Guia simplificado para buscas em bases de patente gratuitas. Módulo 4: Busca no Patentscope**. Diretoria de Patentes, Programas de computador e Topografia de Circuitos Integrados (DIRPA), Coordenação Geral de Estudos, Projetos e Disseminação da Informação Tecnológica (CEPIT), Divisão de Estudos e Projetos (DIESP), 2018. Disponível em: <[https://www.gov.br/inpi/pt-br/assuntos/arquivos-cepit/TUTORIALDEBUSCASModulo4\\_BasePatentscope\\_verso27072018.pdf/view](https://www.gov.br/inpi/pt-br/assuntos/arquivos-cepit/TUTORIALDEBUSCASModulo4_BasePatentscope_verso27072018.pdf/view)>. Acesso em: 19 de dez. de 2023. Citado na página 28.

INPI. **Busca de patentes**. Brasil, 2020. Disponível em: <<https://www.gov.br/inpi/pt-br/assuntos/informacao/busca-de-patentes>>. Acesso em: 05 de jan. de 2020. Citado 2 vezes nas páginas 2 e 19.

INPI. **Convenção de Paris**. Brasil, 2021. Disponível em: <<https://www.gov.br/inpi/pt-br/backup/legislacao-1/cup.pdf>>. Acesso em: 18 de mai. de 2021. Citado na página 19.

INPI. **Indicadores de Propriedade Industrial 2020**. [S.I.], 2021. Disponível em: <[https://www.gov.br/inpi/pt-br/acesso-a-informacao/boletim-mensal/arquivos/documentos/indicadores-2020\\_aecon\\_vf-27-01-2021.pdf](https://www.gov.br/inpi/pt-br/acesso-a-informacao/boletim-mensal/arquivos/documentos/indicadores-2020_aecon_vf-27-01-2021.pdf)>. Citado na página 39.

INPI. **Manual da base de dados sobre propriedade intelectual para fins estatísticos (BADEPI)**. [S.I.], 2021. Disponível em: <[https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/badepi/badepi\\_v8-3\\_patentes.pdf](https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/badepi/badepi_v8-3_patentes.pdf)>. Citado na página 76.

INPI. **Estatísticas e Estudos Econômicos**. Brasil, 2022. Disponível em: <<https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/estatisticas>>. Acesso em: 08 de fev. de 2022. Citado na página 151.

INPI. **Estatísticas**. BRASIL, 2023. Disponível em: <<https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/estatisticas>>. Acesso em: 31 de dez. de 2023. Citado na página 39.

INPI. **Boletim mensal de propriedade industrial: Estatísticas preliminares**. [S.I.], 2024. Disponível em: <[https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-pi\\_resultados-de-setembro-2024\\_v-2.pdf](https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-pi_resultados-de-setembro-2024_v-2.pdf)>. Acesso em: 11 out. 2024. Citado na página 3.

INPI. **Estatísticas**. BRASIL, 2024. Disponível em: <<https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/estatisticas>>. Acesso em: 05 de mai. de 2024. Citado na página 151.

JUNGSMANN, E. A. Diana de M.; BONETTI. **Inovação e Propriedade Intelectual: Guia para o Docente**. [S.I.], 2010. Citado 2 vezes nas páginas 9 e 10.

- JUNIOR, M. R. A. C.; KÜHL. Análise das inovações tecnológicas aplicáveis nas ciências contábeis: Um olhar a partir da bibliometria e patentometria no período 2005-2019. **Brazilian Journal of Development**, v. 6, n. 12, p. 94465–94491, dez. 2020. DOI:10.34117/bjdv6n12-058. Citado na página 43.
- JURGENS, B.; HERRERO-SOLANA, V. Patent bibliometrics and its use for technology watch. **Journal of Intelligence Studies in Business**, v. 7, n. 2, p. 17–26, 2017. Citado na página 11.
- LAKATOS, E. M.; MARCONI, M. de A. **Fundamentos de metodologia científica**. [S.l.]: São Paulo: Atlas, 1993. Citado na página 47.
- LANJOUW, J. O.; SCHANKERMAN, M. Patent quality and research productivity: Measuring innovation with multiple indicators. **The Economic Journal**, n. 114, p. 441–465, 4 2004. Disponível em: <<https://academic.oup.com/ej/article-abstract/114/495/441/5085644?redirectedFrom=fulltext>>. Citado na página 22.
- LATIPAT. **Latipat-Espacenet**. Latipat-Espacenet, 2017. Disponível em: <<https://lp.espacenet.com/>>. Acesso em: 14 de dez. de 2023. Citado na página 27.
- MENA-CHALCO, R. M. J. P.; CESAR-JUNIOR. Scriptlattes: an open-source knowledge extraction system from the lattes platform. **Journal Of The Brazilian Computer Society**, v. 15, n. 4, p. 31–39, dez. 2009. Springer Nature America, Inc <<http://dx.doi.org/10.1007/bf03194511>>. Citado na página 144.
- MENDES, D. R. A. d. M. L. S.; MELO. Avaliação tecnológica: Uma proposta metodológica. **ANPAD - Associação nacional de Pós-graduação e Pesquisa em Administração**, 2017. Disponível em: <<https://doi.org/10.1590/1982-7849rac2017170047>>. Acesso em: 30 jun. de 2021. Citado na página 42.
- MENEZES, M. S. **Estudo patentométrico relacionado à terapia de Tuberculose pulmonar**. Dissertação (Mestrado) — Fundação Oswaldo Cruz, Programa de Pós-graduação Profissional em Gestão, Pesquisa e Desenvolvimento na Indústria Farmacêutica, do Instituto de Tecnologia em Fármacos, 2019. Disponível em: <<https://www.arca.fiocruz.br/handle/icict/37031>>. Citado na página 42.
- MOEHRLE, M. G. et al. Patinformatics as a business process: A guideline through patent research tasks and tools. **World Patent Information**, Elsevier, v. 32, n. 4, p. 291–299, 2010. Disponível em: <<https://ideas.repec.org/a/eee/worpat/v32y2010i4p291-299.html>>. Acesso em: 24 dez. 2023. Citado 2 vezes nas páginas 32 e 33.
- MOURA, A. M. M. de et al. Fontes de informação em patentes: análise das características das bases derwent innovations index, orbit, inpi, google patents e patentscope com base na produção tecnológica da ufrgs. **Folha de Rosto: Revista de Biblioteconomia e Ciência da Informação**, v. 5, n. 2, p. 17–27, 2019. Disponível em: <<https://periodicos.ufca.edu.br/ojs/index.php/folhaderosto/article/view/373/385>>. Citado na página 43.
- MUELLER, S. P. M. Estudos métricos da informação em ciência e tecnologia no brasil realizados sobre a unidade de análise artigos de periódicos. **Liinc em Revista**, p. 6–27, 2013. Citado na página 35.

MUELLER, V. S. P. 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, abr. 2014. Citado 3 vezes nas páginas 1, 12 e 14.

NACKE, O. Informetría: un nuevo nombre para una nueva disciplina. definición, estado de la ciencia y principios de desarrollo. **Revista Española de Documentación Científica**, v. 6, n. 3, jan. 1983. Citado na página 36.

NARIN, F. Patents as indicators for the evaluation of industrial research output. **Scientometrics**, v. 34, n. 3, p. 489–496, jul. 1995. Citado na página 37.

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. Citado 2 vezes nas páginas 1 e 37.

NCET, N. C. E. T. **Management method for classifying patent files by using dates**. 2019. Depositante: Nanjing Chuqing Electronic Technology. CN110309384A. Depósito: 8 set. 2019. Publicação: 8 dez. 2019. Citado na página 2.

OLIVEIRA, L. G. de et al. Informação de patentes: ferramenta indispensável para a pesquisa e o desenvolvimento tecnológico. **Química Nova**, n. 28, p. 36–40, 12 2005. Disponível em: <<https://doi.org/10.1590/S0100-40422005000700007>>. Citado 2 vezes nas páginas 20 e 22.

OPS. **Open Patent Services RESTful Web Services**. 1.3.16. ed. <https://www.epo.org/searching-for-patents/data/web-services/ops.html>, 2023. Citado na página 73.

PARRA, R. X. M. R.; COUTINHO, E. F. C.; PESSANO. Um breve olhar sobre a cienciometria: Origem, evolução, tendências e sua contribuição para o ensino de ciências. **Revista Contexto & Educação**, n. 107, abr. 2019. Disponível em: <<https://www.revistas.unijui.edu.br/index.php/contextoeducacao/article/view/7267>>. Acesso em: 30 mar. de 2021. Citado na página 35.

PIRES, E. A.; RIBEIRO, N. M.; QUINTELLA, C. M. Sistemas de busca de patentes: análise comparativa entre espacenet, patentscope, google patents, lens, derwent innovation index e orbit intelligence. **Cadernos de Prospecção**, v. 13, n. 1, p. 13, mar 2020. Disponível em: <<https://periodicos.ufba.br/index.php/nit/article/view/35147>>. Citado 6 vezes nas páginas 23, 24, 26, 28, 29 e 30.

PORTAL, S. G. Modelo teórico para el estudio métrico de la información documental. **Madrid: TREA**, p. 176, 2005. Citado na página 37.

QUESTEL. **Emerge Smarter With Orbit Intelligence! Meet the Future With Agility & Confidence**. Questel, 2023. Disponível em: <<https://www.questel.com/patent/ip-intelligence-software/orbit-intelligence/>>. Acesso em: 22 de dez. de 2023. Citado na página 30.

QUESTEL. **Patent Database**. Questel, 2025. Disponível em: <<https://www.questel.com/data-scope/>>. Acesso em: 05 de jan. de 2026. Citado na página 30.

QUONIAM, C. T. L.; KNISS, M. R.; MAZIERI. A patente como objeto de pesquisa em ciências da informação e comunicação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 19, n. 39, p. 243–268, abr. 2014. Disponível em: <<https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2014v19n39p243/26586>>. Acesso em: 13 abr. de 2021. Citado 2 vezes nas páginas 9 e 40.

RATUR, M. K. et al. Patinformatics – an emerging scientific discipline. **SSRN**, SSRN, p. 35, 2010. Disponível em: <[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1566067](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1566067)>. Acesso em: 24 dez. 2023. Citado na página 31.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**, 2019. Disponível em: <<http://arxiv.org/abs/1908.10084>>. Citado na página 107.

REYMOND, D.; QUONIAM, L. A new patent processing suite for academic and research purposes. **World Patent Information**, Science Direct, v. 47, p. 40–50, 2016. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/S0172219016301132>>. Citado na página 41.

REYMOND, D.; QUONIAM, L. Patent documents in stem and phd education: Open-source tools and some examples to open discussion. **EDUCON 2018 IEEE Global Engineering Education Conference**, p. 4–9, 2018. Citado na página 2.

REZENDE, N. G. et al. Elaboração de um conjunto de dados sobre o registro de patentes no brasil. In: **DATASET SHOWCASE WORKSHOP (DSW)**. Belo Horizonte: Sociedade Brasileira de Computação, 2023. v. 5, p. 99–108. Disponível em: <<https://sol.sbc.org.br/index.php/dsw/article/view/25509/25329>>. Acesso em: 13 jan. 2024. Citado na página 44.

RIBEIRO, S. P. M. J. C.; GARCIA. O estado da arte da patentometria em periódicos internacionais da ciência da informação. **Encontro Brasileiro de bibliometria e cienciometria**, mai. 2014. Disponível em: <<https://www.brapci.inf.br/index.php/res/v/45400>>. Acesso em: 22 fev. de 2021. Citado na página 41.

RODRIGUES, A. de J. **Metodologia científica: completo e essencial para a vida universitária**. first. [S.l.]: Avercamp, 2006. Citado na página 47.

RUDIO, F. V. **Introdução ao projeto de pesquisa científica**. [S.l.]: Petrópolis: Vozes, 2000. Citado na página 47.

SAESP/DIRPA/INPI, S. de Assuntos Especiais da Diretoria de P. **Manual Para o Depositante de Patentes**. [S.l.], 2015. Disponível em: <[https://www.gov.br/inpi/pt-br/assuntos/arquivos-dirpa/ManualparaoDepositantedePatentes23setembro2015\\_versaoC\\_set\\_15.pdf](https://www.gov.br/inpi/pt-br/assuntos/arquivos-dirpa/ManualparaoDepositantedePatentes23setembro2015_versaoC_set_15.pdf)>. Acesso em: 11 set. de 2021. Citado 2 vezes nas páginas 23 e 188.

SANZ-CASADO, E. **Los estudios métricos de la información y la evaluación del a actividad científica: conceptos básicos**. [S.l.]: São Paulo: ECA/USP, 2006. Material didático de curso “Os estudos métricos da informação”, ministrado no Programa de Pós-Graduação em Ciência da Informação da ECA/USP, novembro 2006. Citado na página 1.

SCARTASSINI, F. V. B. et al. Estudo patentométrico das patentes brasileiras na via pct. **Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (XIX ENANCIB)**, n. XIX ENANCIB, 2018. Disponível em: <<http://hdl.handle.net/20.500.11959/brapci/102164>>. Citado na página 42.

SEBRAE. **Definição de Patente**. Brasil, 2021. Disponível em: <<https://www.sebrae.com.br/sites/PortalSebrae/artigos/definicao-de-patente,230a634e2ca62410VgnVCM100000b272010aRCRD>>. Acesso em: 13 de jan. de 2021. Citado na página 11.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. fourth. [S.l.]: Universidade Federal de Santa Catarina - UFSC, 2005. Disponível em: <[https://tccbiblio.paginas.ufsc.br/files/2010/09/024\\_Metodologia\\_de\\_pesquisa\\_e\\_elaboracao\\_de\\_teses\\_e\\_dissertacoes1.pdf](https://tccbiblio.paginas.ufsc.br/files/2010/09/024_Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes1.pdf)>. Acesso em: 15 out. 2024. Citado na página 47.

SILVA, M. d. L. P. José Aparecido da; BIANCHI. Cientometria: a métrica da ciência. **Paidéia (Ribeirão Preto)**, v. 11, n. 21, p. 5–10, 2001. Citado na página 34.

SILVA, M. d. L. P. José Aparecido da; BIANCHI. Cientometria: a métrica da ciência. **Paidéia (Ribeirão Preto)**, v. 11, n. 21, 2001. Disponível em: <<https://doi.org/10.1590/S0103-863X2001000200002>>. Acesso em: 29 mar. de 2021. Citado 2 vezes nas páginas 35 e 36.

SILVA, M. G. S. R. da; NASCIMENTO. Patentometria: uma ferramenta indispensável no estudo de desenvolvimento de tecnologias para a indústria química. **Quim. Nova**, v. 43, n. 10, p. 1538–1548, ago. 2020. DOI: <<http://dx.doi.org/10.21577/0100-4042.20170620>>. Acesso em: 08 dez. de 2021. Citado 3 vezes nas páginas 2, 38 e 44.

SILVA, R. R. da. **Coleta e tratamento de dados sobre a produção técnica brasileira: Um estudo baseado em patentes**. Dissertação (Mestrado) — Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-graduação em Modelagem Matemática e Computacional, 2022. Disponível em: <<https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=4083005&key=cd664e534f652c7182bf8597951e72fd>>. Citado na página 33.

SINGH, V.; CHAKRABORTY, K.; LAVINA-VINCENT, C. Patent database: Their importance in prior art documentation and patent search. **Journal of Intellectual Property Rights**, n. 21, 1 2016. Citado na página 31.

SOUZA, A. A. A. S. N. et al. Mapeamento de depósitos de propriedade intelectual no inpi por discentes do programa de pós-graduação em ciência da propriedade intelectual da universidade federal de sergipe. **11TH INTERNATIONAL SYMPOSIUM ON TECHNOLOGICAL INNOVATION**, v. 11, n. 1, p. 1524–1533, apr 2021. D.O.I.: 10.51722/S2318-3403202100011344, ISSN: 2318-3403. Citado na página 44.

SÁNCHEZ, M. V. G. **Patentometría: herramienta para el análisis de oportunidades tecnológicas**. Tese (Doutorado em Gerência de Información Tecnológica) — Facultad de Economía - Universidad de La Habana, La Habana, dez. 1999. Citado na página 37.

TAGUE-SUTCLIFFE, J. An introduction to informetrics. **Information Processing & Management**, v. 28, p. 1–3, 1992. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/030645739290087G>>. Acesso em: 16 abr. de 2021. Citado 2 vezes nas páginas 35 e 37.

TAUBES, G. Measures for measure in science. **Science**, n. 260, p. 884–886, 1993. Citado na página 36.

TELES, A. L. et al. Desenvolvimento de um sistema para gestão de propriedade intelectual para instituições de ciência e tecnologia. In: **Demonstrações e Aplicações - Simpósio Brasileiro de Banco de Dados (SBBDD)**. Belo Horizonte: Sociedade Brasileira de Computação, 2023. v. 38, p. 132–137. Disponível em: <[https://sol.sbc.org.br/index.php/sbbdd\\_estendido/article/view/25626](https://sol.sbc.org.br/index.php/sbbdd_estendido/article/view/25626)>. Acesso em: 13 jan. 2024. Citado na página 45.

THE LENS. **About The Lens**. Cambia, 2023. Disponível em: <<https://about.lens.org/#>>. Acesso em: 19 de dez. de 2023. Citado na página 29.

THE SQLITE CONSORTIUM. **THE SQLITE CONSORTIUM**. 2023. Disponível em: <<https://sqlite.org/index.html>>. Acesso em: 22 de mar. de 2023. Citado na página 88.

TRIPPE, A. **Guidelines for Preparing Patent Landscape Reports**. [S.l.]: World Intellectual Property Organization (WIPO), 2015. Disponível em: <<https://www.wipo.int/publications/en/details.jsp?id=3938&plang=EN>>. Acesso em: 07 mai. de 2021. Citado na página 38.

UECE. **INPI - Saiba mais sobre a nova numeração nos pedidos da DIRPA e da DICIG**. Universidade Federal do Ceará, 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](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)>. Acesso em: 11 de mai. de 2021. Citado na página 99.

USPTO. **Overview of the U.S. Patent Classification System (USPC)**. USPTO, 2012. Disponível em: <<https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf>>. Acesso em: 24 de nov. de 2023. Citado na página 18.

VANTI, N. Os links e os estudos webométricos. **Ciência da Informação**, v. 34, n. 01, 2005. Disponível em: <<https://doi.org/10.1590/S0100-19652005000100009>>. Acesso em: 09 abr. de 2021. Citado na página 36.

VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, v. 31, n. 02, p. 152–162, 2002. Disponível em: <<https://doi.org/10.1590/S0100-19652002000200016>>. Acesso em: 09 abr. de 2021. Citado 2 vezes nas páginas 35 e 36.

VELHO, L. Indicadores científicos: em busca de teoria. **interciência**. **Caracas**, v. 15, n. 3, p. 139–145, 1990. Citado na página 35.

WIPO. **WIPO Patent Drafting Manual**. <https://www.wipo.int/publications/en/details.jsp?id=297>, 2007. Citado na página 19.

WIPO. **STANDARD ST.9**. WIPO, 2013. Disponível em: <<https://www.wipo.int/export/sites/www/standards/en/pdf/03-09-01.pdf>>. Acesso em: 23 de nov. de 2023. Citado na página 22.

WIPO. **Sistema de Classificação de patentes**. 2020. Disponível em: <<http://ipc.inpi.gov.br/classifications/ipc/ipcpub/?notion=scheme&version=20200101&symbol=none&menulang=pt&lang=pt&viewmode=f&fipcp=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart>>. Acesso em: 15 de mar. de 2020. Citado na página 15.

WIPO. **Família de Patentes**. Brasil, 2021. Disponível em: <[https://www.wipo.int/edocs/mdocs/aspac/en/wipo\\_ip\\_bkk\\_12/wipo\\_ip\\_bkk\\_12\\_www\\_238983.pdf](https://www.wipo.int/edocs/mdocs/aspac/en/wipo_ip_bkk_12/wipo_ip_bkk_12_www_238983.pdf)>. Acesso em: 20 de abr. de 2021. Citado na página 19.

WIPO. **PCT – O Sistema Internacional de Patentes**. WIPO, 2023. Disponível em: <<https://www.wipo.int/pct/pt/>>. Acesso em: 23 de nov. de 2023. Citado 2 vezes nas páginas 19 e 20.

WIPO, W. I. P. O. **Com crescimento impulsionado pela Ásia, depósitos de PI em todo o mundo registram novo recorde histórico em 2021**. WIPO, 2022. Disponível em: <[https://www.wipo.int/pressroom/pt/articles/2022/article\\_0013.html](https://www.wipo.int/pressroom/pt/articles/2022/article_0013.html)>. Acesso em: 11 de out. de 2024. Citado na página 3.

YUE, L. B. et al. Social network information visualization based on fruchterman reingold layout algorithm. **2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)**, mar 2018. <<https://ieeexplore.ieee.org/abstract/document/8367690>>. Acesso em: 08 set. de 2023. Citado na página 157.

## Apêndices

# APÊNDICE A – Resultados da busca no Google

Tabela 18 – Resultados considerados na análise

| nº | Página da Web                                                | Endereço eletrônico                                                                                                                                                                         |
|----|--------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 01 | Patent search — Instituto Nacional da Propriedade Industrial | <a href="https://www.gov.br/inpi/en/services/patents/basic-guide/patent-search">https://www.gov.br/inpi/en/services/patents/basic-guide/patent-search</a>                                   |
| 02 | Search - Patent Center - USPTO                               | <a href="https://patentcenter.uspto.gov/search">https://patentcenter.uspto.gov/search</a>                                                                                                   |
| 03 | The Lens                                                     | <a href="https://support.lens.org/knowledge-base/patent-search/">https://support.lens.org/knowledge-base/patent-search/</a>                                                                 |
| 04 | INPI - Instituto Nacional da Propriedade Industrial          | <a href="https://www.gov.br/inpi/en">https://www.gov.br/inpi/en</a>                                                                                                                         |
| 05 | Patents                                                      | <a href="https://www.uspto.gov/patents">https://www.uspto.gov/patents</a>                                                                                                                   |
| 06 | Patent Search / Utility model Search   Japan Patent Office   | <a href="https://www.jpo.go.jp/e/support/j_platpat/patent_search.htm">https://www.jpo.go.jp/e/support/j_platpat/patent_search.htm</a>                                                       |
| 07 | Google Patents                                               | <a href="https://patents.google.com/?scholar">https://patents.google.com/?scholar</a>                                                                                                       |
| 08 | INPI - patent search                                         | <a href="https://www.inpi.pt/services/patent_search.html">https://www.inpi.pt/services/patent_search.html</a>                                                                               |
| 09 | Patent Search                                                | <a href="https://www.patsnap.com/glossary/patent-search/">https://www.patsnap.com/glossary/patent-search/</a>                                                                               |
| 10 | Patents   Intellectual Property India   Government of India  | <a href="https://ipindia.gov.in/patents.htm">https://ipindia.gov.in/patents.htm</a>                                                                                                         |
| 11 | Patents Search - IPOI                                        | <a href="https://www.ipoi.gov.ie/en/ip-search-tools/patents-search/">https://www.ipoi.gov.ie/en/ip-search-tools/patents-search/</a>                                                         |
| 12 | Patent Search Service                                        | <a href="https://www.questel.com/patent/patent-services/patent-search/">https://www.questel.com/patent/patent-services/patent-search/</a>                                                   |
| 13 | PQAI - Patent Search Tool                                    | <a href="https://search.projectpq.ai/">https://search.projectpq.ai/</a>                                                                                                                     |
| 14 | CAS PatentPak® - Full-Text Patent Search                     | <a href="https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/patentpak">https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/patentpak</a> |
| 15 | Patentfield   AI Patent Search, Analytics and investigating  | <a href="https://en.patentfield.com/">https://en.patentfield.com/</a>                                                                                                                       |
| 16 | Simple search engine                                         | <a href="https://ewyszukiwarka.pue.uprp.gov.pl/">https://ewyszukiwarka.pue.uprp.gov.pl/</a>                                                                                                 |
| 17 | Search - BLM GLO Records - Bureau of Land Management         | <a href="https://glorerecords.blm.gov/search/default.aspx">https://glorerecords.blm.gov/search/default.aspx</a>                                                                             |
| 18 | Lens Patent Search                                           | <a href="https://www.lens.org/lens/search/patent/list">https://www.lens.org/lens/search/patent/list</a>                                                                                     |

Continuação da Tabela 18

| nº | Página da Web                                                | Endereço eletrônico                                                                                                                                                                                                                       |
|----|--------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 19 | Expert Patent Search & Opinions                              | < <a href="https://www.dennemeyer.com/patent-protection/patent-searches-opinions/">https://www.dennemeyer.com/patent-protection/patent-searches-opinions/</a> >                                                                           |
| 20 | NLPatent                                                     | < <a href="https://www.nlpatent.com/">https://www.nlpatent.com/</a> >                                                                                                                                                                     |
| 21 | The Global Patent Search Tool for IP Users of All Levels     | < <a href="https://www.inquartik.com/patentcloud/patent-search/">https://www.inquartik.com/patentcloud/patent-search/</a> >                                                                                                               |
| 22 | What is a Patent Search?                                     | < <a href="https://www.oppenhuizen.com/what-is-a-patent-search/">https://www.oppenhuizen.com/what-is-a-patent-search/</a> >                                                                                                               |
| 23 | Search Patent – The Official Portal of Intellectual Property | < <a href="https://www.myipo.gov.my/en/search-patent/">https://www.myipo.gov.my/en/search-patent/</a> >                                                                                                                                   |
| 24 | Patent Search Services   Intellectual Property Strategy      | < <a href="https://www.anaqua.com/services/patent-search-services/">https://www.anaqua.com/services/patent-search-services/</a> >                                                                                                         |
| 25 | Assisted Patent Search - Swiss Federal Institute             | < <a href="https://www.ige.ch/en/services/searches/patent-searches-in-general/assisted-searches/assisted-patent-search">https://www.ige.ch/en/services/searches/patent-searches-in-general/assisted-searches/assisted-patent-search</a> > |
| 26 | Searching                                                    | < <a href="https://support.google.com/faqs/answer/7049475?hl=en">https://support.google.com/faqs/answer/7049475?hl=en</a> >                                                                                                               |
| 27 | Patsnap   Data, Analytics and Expertise for Innovative       | < <a href="https://www.patsnap.com/">https://www.patsnap.com/</a> >                                                                                                                                                                       |
| 28 | Patent Search Services                                       | < <a href="https://clarivate.com/products/patent-intelligence/patent-search-services/">https://clarivate.com/products/patent-intelligence/patent-search-services/</a> >                                                                   |
| 29 | Patent search and trend analysis                             | < <a href="https://ieeexplore.ieee.org/document/7154759">https://ieeexplore.ieee.org/document/7154759</a> >                                                                                                                               |
| 30 | Intellectual Property India                                  | < <a href="https://iprsearch.ipindia.gov.in/PublicSearch/PublicationSearch/ApplicationStatus">https://iprsearch.ipindia.gov.in/PublicSearch/PublicationSearch/ApplicationStatus</a> >                                                     |
| 31 | How to Do a Patent Search in 5 Steps: What You Need          | < <a href="https://www.inquartik.com/blog/basic-how-to-do-a-patent-search-in-5-steps">https://www.inquartik.com/blog/basic-how-to-do-a-patent-search-in-5-steps</a> >                                                                     |
| 32 | Patent Searching Software - Orbit Express                    | < <a href="https://www.questel.com/patent/ip-intelligence-software/orbit-express/">https://www.questel.com/patent/ip-intelligence-software/orbit-express/</a> >                                                                           |
| 33 | Search   Monaco Industrial Property Office                   | < <a href="https://mcipo.gouv.mc/en/content/patent/search">https://mcipo.gouv.mc/en/content/patent/search</a> >                                                                                                                           |
| 34 | Patent Search                                                | < <a href="https://www.rws.com/intellectual-property-solutions/resources/patent-search-brochure/">https://www.rws.com/intellectual-property-solutions/resources/patent-search-brochure/</a> >                                             |
| 35 | Patent Register Search - Swiss Federal Institute             | < <a href="https://www.ige.ch/en/services/searches/patent-searches-in-general/patent-register-search">https://www.ige.ch/en/services/searches/patent-searches-in-general/patent-register-search</a> >                                     |
| 36 | Patent Search 101: Why a US Patent Search is Critically      | < <a href="https://ipwatchdog.com/2018/01/13/patent-search-101-patent-searches">https://ipwatchdog.com/2018/01/13/patent-search-101-patent-searches</a> >                                                                                 |

Continuação da Tabela 18

| nº | Página da Web                                            | Endereço eletrônico                                                                                                                                                                                               |
|----|----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 37 | Search and Enquiry                                       | < <a href="https://digitalhub.ipos.gov.sg/FAMN/eservice/IP4SG/MN_BasicSearch">https://digitalhub.ipos.gov.sg/FAMN/eservice/IP4SG/MN_BasicSearch</a> >                                                             |
| 38 | Powerful new search tool will help IPO maintain patent   | < <a href="https://www.gov.uk/government/news/">https://www.gov.uk/government/news/</a> >                                                                                                                         |
| 39 | How AI Can Improve Patent Search & Monitoring            | < <a href="https://iamip.com/how-ai-can-improve-patent-search-monitoring/">https://iamip.com/how-ai-can-improve-patent-search-monitoring/</a> >                                                                   |
| 40 | SciFinder Training Videos - Patent Searching             | < <a href="https://www.cas.org/support/training/scifinder/patent-searching">https://www.cas.org/support/training/scifinder/patent-searching</a> >                                                                 |
| 41 | How to Do a Patent Search on Google - YouTube            | < <a href="https://www.youtube.com/watch?v=NQnUVr67JK4">https://www.youtube.com/watch?v=NQnUVr67JK4</a> >                                                                                                         |
| 42 | Derwent Innovation - Patent Search for IP Professionals  | < <a href="https://clarivate.com/products/ip-intelligence/patent-intelligence-software/derwent-innovation/">https://clarivate.com/products/ip-intelligence/patent-intelligence-software/derwent-innovation/</a> > |
| 43 | Apply for a patent: Search, publication and 'substantive | < <a href="https://www.gov.uk/patent-your-invention/request-your-search-and-examination">https://www.gov.uk/patent-your-invention/request-your-search-and-examination</a> >                                       |
| 44 | Similarity search in patents databases.                  | < <a href="https://www.sciencedirect.com/science/article/pii/S0172219021000041">https://www.sciencedirect.com/science/article/pii/S0172219021000041</a> >                                                         |
| 45 | rbouadjenek/patent-search                                | < <a href="https://github.com/rbouadjenek/patent-search">https://github.com/rbouadjenek/patent-search</a> >                                                                                                       |
| 46 | 5 Step Patent Search using ChatGPT - YouTube             | < <a href="https://www.youtube.com/watch?v=JI9o5fR3kQk">https://www.youtube.com/watch?v=JI9o5fR3kQk</a> >                                                                                                         |
| 47 | Guide to searching of the patent register                | < <a href="https://www.iponz.govt.nz/about-ip/patents/search/searching-guide/">https://www.iponz.govt.nz/about-ip/patents/search/searching-guide/</a> >                                                           |
| 48 | Taiwan Intellectual Property Office-Services-Search      | < <a href="https://www.tipo.gov.tw/en/lp-277-2.html">https://www.tipo.gov.tw/en/lp-277-2.html</a> >                                                                                                               |
| 49 | Assisted patent searches                                 | < <a href="https://www.innosuisse.ch/patent">https://www.innosuisse.ch/patent</a> >                                                                                                                               |
| 50 | InnovationQ+ Patent Search Software & Analytics Platform | < <a href="https://ip.com/innovation-power-suite/innovation-q-patent-search-software">https://ip.com/innovation-power-suite/innovation-q-patent-search-software</a> >                                             |

## APÊNDICE B – Elementos da estrutura do número de depósito de patentes

Quadro 19 – Códigos de natureza de proteção que compõem o número de depósito da patente

| 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                                     |

Fonte: adaptado de [SAESP/DIRPA/INPI \(2015\)](#).

Quadro 20 – Códigos de publicação que compõem o número de depósito da patente

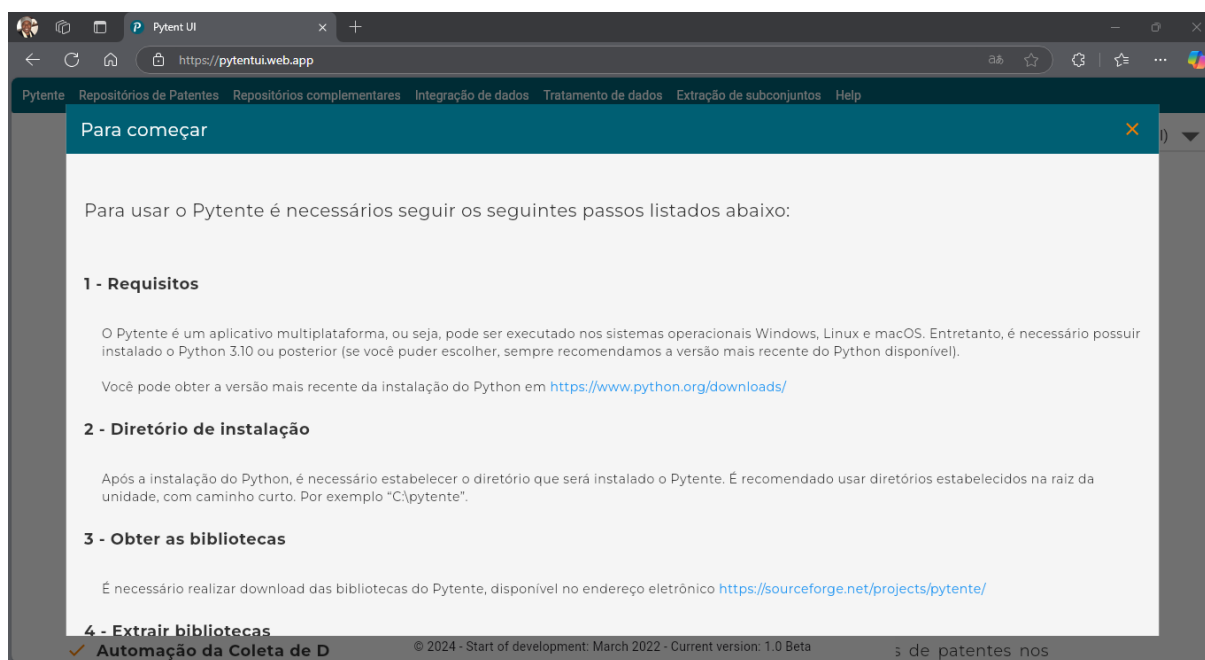
| Código | Descrição                                     |
|--------|-----------------------------------------------|
| A1     | Publicação do pedido de patente               |
| A2     | Publicação do pedido sem o relatório de busca |
| A3     | Publicação do pedido com o relatório de busca |
| B1 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

Fonte: adaptado de [SAESP/DIRPA/INPI \(2015\)](#).

## APÊNDICE C – Instalação e configuração do Pytente

A instalação do Pytente foi desenvolvida para ser acessível a usuários com diferentes níveis de familiaridade com recursos computacionais, abrangendo desde aqueles com pouca experiência até usuários avançados. Para facilitar esse processo, a interface web disponibiliza um guia passo a passo acessível no menu [Pytente > Por onde começar], com orientações detalhadas sobre cada etapa do processo de instalação [Figura 59](#).

Figura 59 – Página do guia passo a passo de instalação do Pytente



Fonte: Elaboração do autor

Conforme apresentado na figura, o Pytente disponibiliza uma página específica para auxiliar o pesquisador no processo de instalação e configuração da ferramenta. Esse processo é estruturado em sete etapas, abrangendo desde a instalação das dependências até a realização de testes para verificar a conclusão bem sucedida da instalação.

A primeira etapa consiste na instalação das dependências necessárias. Nesta versão inicial, a única dependência requerida é o Python, sendo necessária a instalação da versão 3.10 ou superior. Concluída essa etapa, o segundo passo envolve a definição do diretório de instalação do Pytente. É recomendado que esse diretório seja criado diretamente na raiz da unidade de armazenamento, a fim de evitar caminhos excessivamente longos, como, por

exemplo, "C: pytente".

No terceiro passo, é necessário realizar o *download* das bibliotecas do Pytente, disponíveis no repositório eletrônico <<https://sourceforge.net/projects/pytente/>>. Essas bibliotecas estão agrupadas em um único arquivo compactado. Após a conclusão do *download*, o quarto passo consiste na extração de todo o conteúdo do arquivo para o diretório previamente definido na segunda etapa.

Com todos os recursos necessários disponíveis, inicia-se a fase de configuração. Assim, no quinto passo, é necessário criar um ambiente virtual para evitar conflitos entre as bibliotecas do Pytente e outras ferramentas instaladas no computador. Para isso, é necessário acessar o diretório de instalação do Pytente, por meio do "Prompt de Comando" (Windows) ou do "Terminal" (Linux e macOS). Em seguida, executar o seguinte comando para criar o ambiente virtual:

```
python -m venv venv
```

Após a criação do ambiente virtual, é necessário ativá-lo. Para sistemas Windows, utilize:

```
.\venv\Scripts\Activate
```

Já para sistemas Linux e macOS, utilize:

```
source venv/bin/activate
```

No sexto passo, é necessário instalar as bibliotecas requeridas pelo Pytente executando o comando:

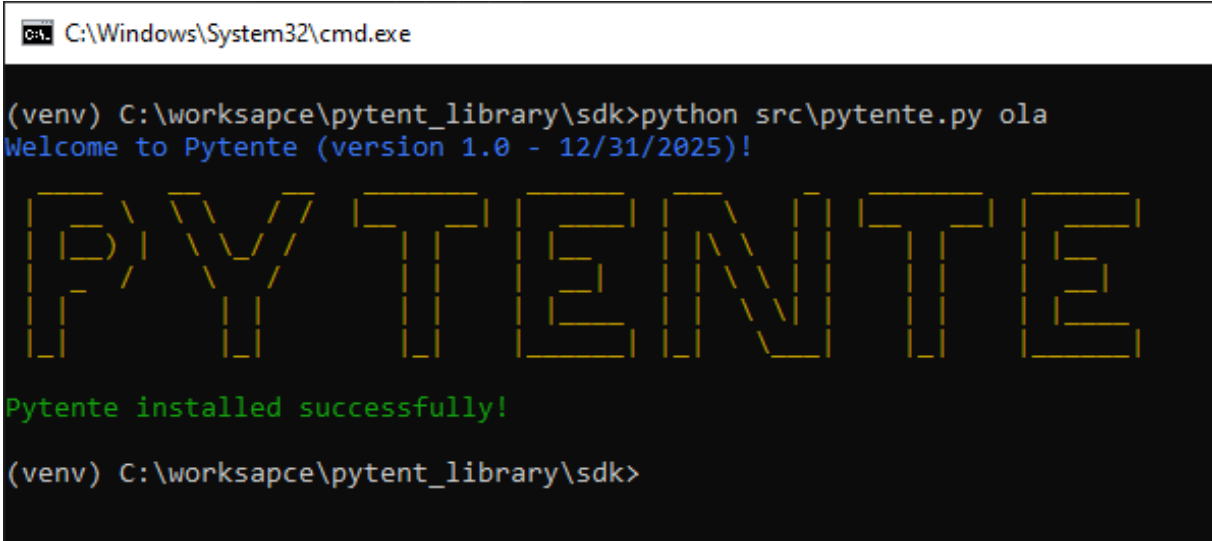
```
pip install -r requirements.txt
```

Por fim, no sétimo passo, é necessário verificar se a instalação foi bem sucedida. Para isso, é necessário executar o comando:

```
python src\pytente.py ola
```

Caso o processo tenha sido concluído corretamente, a saída gerada no terminal deverá ser semelhante à apresentada na [Figura 60](#).

Figura 60 – Página do guia passo a passo de instalação do Pytente



```
C:\Windows\System32\cmd.exe

(venv) C:\workspace\pytent_library\sdk>python src\pytente.py ola
Welcome to Pytente (version 1.0 - 12/31/2025)!

PYTENTE

Pytente installed successfully!

(venv) C:\workspace\pytent_library\sdk>
```

Fonte: Elaboração do autor

Se o terminal exibir uma saída semelhante à apresentada na figura, isso indica que o Pytente foi instalado corretamente e está pronto para ser utilizado nos processos de coleta, tratamento e extração de subconjuntos de dados.

## C.1 Coleta de dados bibliográficos de patentes

Uma das principais funcionalidades do Pytente é a automação do processo de coleta de grandes volumes de dados de patentes em repositórios de acesso aberto, como a Espacenet. Por meio do menu [Repositórios de Patentes > Espacenet], o usuário pode acessar a página destinada à definição da estratégia de coleta de dados bibliográficos no repositório da Espacenet ([Figura 61](#)). No mesmo menu [Repositórios de Patentes] é possível ter acesso a coleta em outros repositórios de patentes como o Google Patents.

Figura 61 – Página de coleta de dados no repositório Espacenet

The screenshot shows the 'Coletar patentes na Espacenet' configuration page in the Pytente UI. The page has a dark teal header with the title and a close button. Below the header, there's a section titled 'Configuração da coleta de patentes na Espacenet'. It contains several input fields: 'COSUMER KEY' and 'CONSUMER SECRET KEY' (note the typo in the image), a 'Contextual Query Language (CQL)' field with a sample query 'pd="20240101 20240102" AND pn="BR" AND ipc="A01C"', and a 'Diretório de armazenamento dos resultados de coleta' field with the path 'C:\patentes\espacenet'. There's a 'Download do arquivo de configuração' button. At the bottom, there are instructions for Windows and Linux/Mac users, and a footer with copyright information and version details.

Fonte: Elaboração do autor

Conforme apresentado na figura, a coleta de dados no repositório Espacenet exige o fornecimento de algumas informações obrigatórias. Primeiramente, é necessário inserir as credenciais de identificação da plataforma, denominadas "Consumer Key" e "Consumer Secret Key". Essas credenciais são pessoais e intransferíveis, sendo necessário que cada pesquisador interessado em recuperar dados na Espacenet possua seu próprio acesso, conforme exigência da plataforma. Para obter, é necessário realizar um cadastro no site da Espacenet, através da página <https://www.epo.org/en/searching-for-patents/data/web-services/ops>.

Outro elemento fundamental é a consulta CQL, que define os critérios de seleção das patentes a serem recuperadas. A CQL (*Common Query Language*) é uma linguagem formal adotada pela Espacenet para estruturar buscas em seu catálogo bibliográfico de patentes. Por exemplo, para recuperar patentes depositadas entre 01/01/2024 e 31/01/2024, depositadas no Brasil e classificadas nas áreas de plantio, semeadura e fertilizantes, deve-se utilizar a seguinte consulta:

```
pd="20240101 20240131" AND pn="BR" AND ipc="A01C"
```

Para mais informações sobre a construção de consultas CQL, é necessário consultar a documentação oficial da Espacenet.

Após a definição da consulta, é necessário especificar o diretório local onde os arquivos recuperados do repositório Espacenet serão armazenados. Concluída essa etapa,

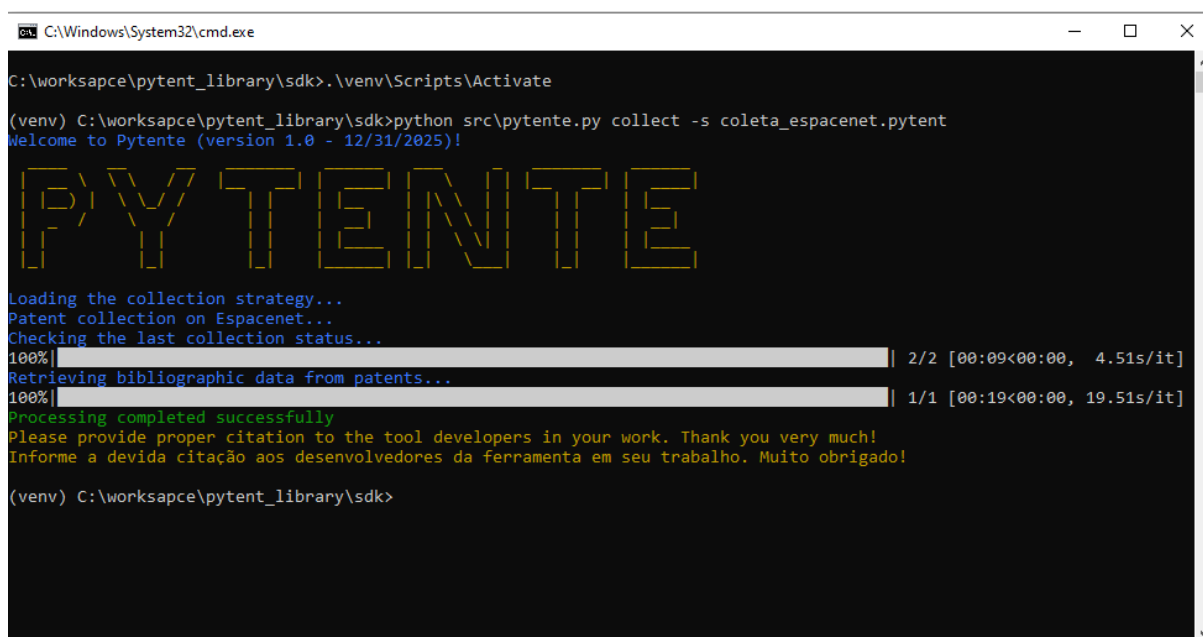
deve-se realizar o *download* do arquivo (coleta\_espacenet.pytent) de configuração e movê-lo para a pasta [input] dentro do diretório de instalação do Pytente.

Com essa etapa finalizada, o próximo passo consiste em acessar o diretório de instalação do Pytente utilizando o Prompt de Comando (Windows) ou Terminal (Linux/Mac) e ativar o ambiente virtual do Pytente. Feito isso, basta executar o seguinte comando para iniciar a coleta de dados:

```
python src\pytente.py collect -s coleta_espacenet.pytent
```

Esse comando dará início ao processo de busca e extração dos dados bibliográficos das patentes. Caso a operação seja bem-sucedida, uma tela semelhante à apresentada na [Figura 62](#) será exibida.

Figura 62 – Tela de sucesso de execução do processo de coleta de dados bibliográficos de patentes no repositório Espacenet



Fonte: Elaboração do autor

É relevante destacar que o tempo necessário para a realização da coleta pode variar conforme o volume de patentes que atendam aos critérios de busca estabelecidos. No entanto, caso haja necessidade de interromper o processo, este pode ser finalizado e posteriormente retomado a partir do ponto onde foi interrompido, bastando, para isso, executar novamente o comando de coleta.

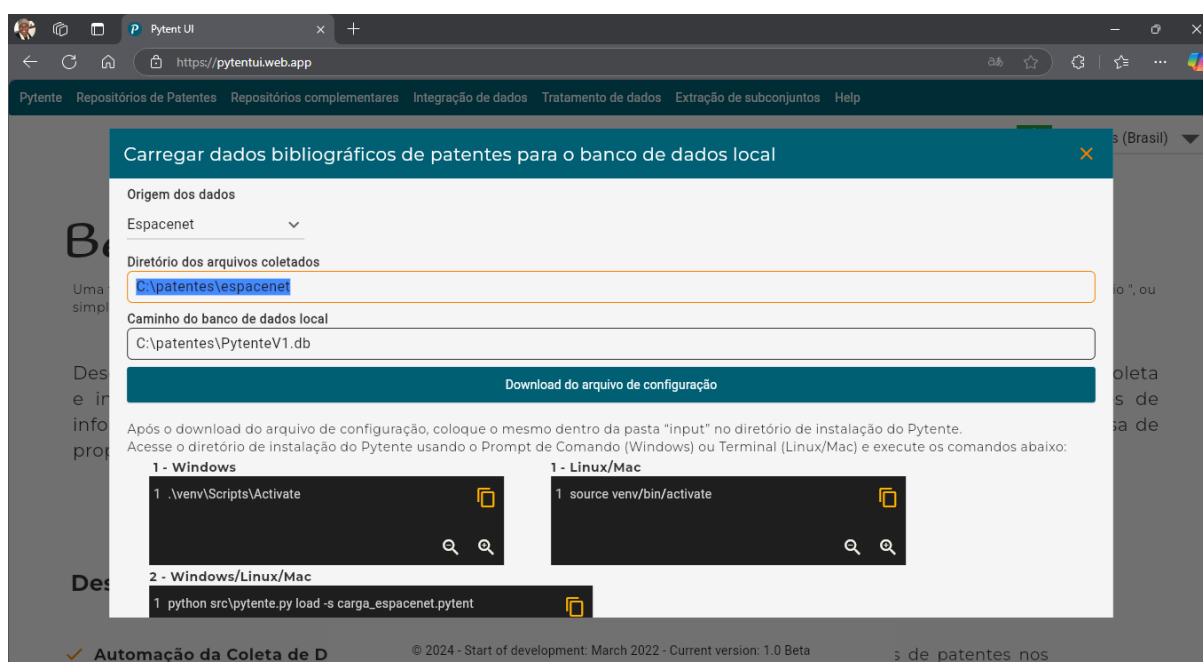
Os dados coletados pelo Pytente são armazenados em arquivos com extensão `.json` no diretório especificado no arquivo de estratégia de coleta. Esse formato de armazenamento possibilita a utilização dos dados em diversas ferramentas de tratamento

e integração, conferindo maior flexibilidade ao pesquisador na análise e manipulação das informações.

## C.2 Tratamento e integração de dados

O Pytente também oferece ferramentas para o tratamento e a integração dos dados coletados, permitindo a construção de um banco de dados local. Por meio do menu [Integração de dados > Download do banco de dados (vazio)], é possível realizar o *download* de um banco de dados relacional vazio, que será utilizado pelo Pytente para armazenar os dados coletados. Feito o *download* do banco de dados vazio, no menu [Integração de dados > Carga de dados bibliográficos de patentes], é possível configurar o carregamento dos dados coletados (Figura 63).

Figura 63 – Tela de configuração de carga de dados bibliográficos de patentes



Fonte: Elaboração do autor

Essa tela permite a definição de todos os parâmetros necessários para importar o conteúdo dos arquivos e armazená-lo no banco de dados local. Inicialmente, é necessário informar a origem dos dados coletados, pois a estrutura dos arquivos pode variar conforme a fonte de coleta. Em seguida, é necessário informar o diretório onde os arquivos coletados estão armazenados, bem como o caminho do banco de dados vazio.

Após o preenchimento dessas informações, o próximo passo consiste em realizar o *download* do arquivo de configuração (*carga\_espacenet.pytent*) e armazená-lo na pasta [input] do diretório de instalação do Pytente. Assim como no processo de coleta de dados, é



Esse comando processa a base de dados, identificando e removendo registros inválidos ou inconsistentes, com base nos critérios estabelecidos pelo Protocolo CTDP.

Concluída a etapa de validação, a deduplicação dos dados pode ser configurada por meio do menu [Tratamento de dados > Deduplicação de dados]. Para esse procedimento, é necessário informar o caminho do banco de dados local e copiar o arquivo de configuração correspondente (deduplicate.pytent) para a pasta [input]. Em seguida, executar o comando:

```
python src\pytente.py deduplicate -s deduplicate.pytent
```

Esse comando elimina registros duplicados da base de dados, aplicando os critérios de deduplicação definidos pelo Protocolo CTDP.

Após a conclusão das etapas de carga, validação e deduplicação, o banco de dados local estará estruturado e pronto para a realização de análises e extração de subconjuntos.

### C.3 Extração de subconjuntos

Para a extração de subconjuntos de dados, o Pytente recomenda a utilização do *DB Browser for SQLite* (DB4S), um navegador de banco de dados que pode ser acessado por meio do menu [Integração de dados > Navegador de banco de dados]. O DB4S é uma ferramenta de software livre projetada para usuários que necessitam criar, pesquisar e editar arquivos de banco de dados SQLite. Além de oferecer uma interface semelhante a uma planilha, a ferramenta disponibiliza suporte para consultas SQL, sendo compatível com sistemas operacionais Windows, macOS e a maioria das distribuições Linux e Unix.

O uso do DB4S representa a abordagem mais flexível para a recuperação de dados, uma vez que permite o acesso direto à estrutura do banco de dados, incluindo nomes de tabelas e campos. Dessa forma, os pesquisadores podem construir consultas SQL personalizadas conforme suas necessidades específicas de análise.

Entretanto, para facilitar o acesso aos dados por pesquisadores, que não possuem familiaridade com a construção de consultas SQL, o Pytente também disponibiliza a extração de subconjuntos predefinidos, os quais serão abordados na [Seção 7.1](#). Esses subconjuntos podem ser acessados diretamente, por meio do menu [Extração de subconjuntos > Modelos pré-definidos].