Elaboração de um Conjunto de Dados sobre o Registro de Patentes no Brasil

Nicolas G. Rezende[1], **Daniel Hasan Dalip**[1],
Michele A. Brandão[2,3], **Marisa A. Vasconcelos**[4]

[1]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Belo Horizonte, MG – Brasil

[2]Instituto Federal de Minas Gerais (IFMG)
Ribeirão das Neves, MG – Brasil

[3]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brasil

[4]IBM Research
São Paulo, SP - Brasil

nicolas.gomes.rezende@gmail.com, hasan@cefetmg.br,
michele.brandao@dcc.ufmg.br, marisaav@br.ibm.com

Abstract. *This paper presents a dataset developed from the information available in the patent registration journals of the Instituto Nacional de Propriedade Industrial (INPI). Due to the fragmentation of information across multiple journals, it becomes impractical to perform searches or conduct analyses on the patent registration landscape in Brazil. The dataset presented in this article aggregates the fragmented information from the processes into a repository on GitHub and employs an automated pipeline using GitHub Actions to keep the dataset updated and relevant.*

Resumo. *Este artigo apresenta um conjunto de dados desenvolvido a partir das informações disponibilizadas nas revistas de registro de patentes do Instituto Nacional de Propriedade Industrial (INPI). Devido à fragmentação das informações em múltiplas revistas, torna-se inviável realizar buscas ou fazer análises sobre o cenário de registro de patentes no Brasil. O conjunto de dados apresentado neste artigo, agrega as informações fragmentadas dos processos em um repositório no Github e apresenta um fluxo automatizado para obtenção de novas revistas, utilizando Github Actions com o objetivo de manter os dados atualizados e relevantes.*

1. Introdução

O Instituto Nacional da Propriedade Industrial (INPI) foi criado em 1970 com o objetivo de regulamentar e proteger registros de marcas, de programas de computador, de patentes, de desenhos industriais, de contratos de tecnologia e de indicações geográficas. Para manter o público informado e garantir a transparência do sistema, o INPI disponibiliza, semanalmente, a Revista de Propriedade Intelectual (RPI)[1], na qual detalha os eventos e as exigências relacionadas aos processos de proteção em andamento.

[1]Revista de Propriedade Intelectual: http://revistas.inpi.gov.br/

A estrutura dos dados de uma RPI apresenta diversas limitações que dificultam sua utilização. As informações sobre os processos estão dispersas em várias revistas, o que torna a busca por dados uma tarefa trabalhosa e desmotivadora. Consequentemente, até mesmo análises simples, como contar o número de registros de patentes em um determinado ano, exigem um profundo conhecimento da estrutura dos dados e a implementação de um algoritmo para processar centenas de revistas. Por esse motivo, nota-se que os dados disponibilizados nas revistas, embora ricos em informações e com grande potencial para várias aplicações, são pouco explorados por estudantes e pesquisadores.

A única alternativa à Revista de Propriedade Industrial é o sistema BuscaWeb[2] disponibilizado pelo INPI. Apesar de apresentar os dados de forma mais amigável, a plataforma não é uma boa fonte de dados para realizar pesquisas, pois frequentemente encontra-se indisponível, não possui uma API pública e oferece opções limitadas de busca. Além disso, nem mesmo o próprio INPI considera o sistema como uma fonte oficial dos dados e recomenda a utilização somente das revistas para a consulta [INPI 2020].

Visando minimizar esses problemas, este trabalho apresenta duas contribuições: um conjunto de dados inédito para comunidade científica que consiste em informações sobre o registro de patentes desde a edição 2474 da RPI publicada em junho de 2018 e um processo que automatiza a obtenção e processamento das revistas. As informações sobre cada processo de proteção de Propriedade Intelectual (PI) foram agrupadas em arquivos separados e, em cada um, é possível encontrar todas as informações já publicadas sobre uma PI nas revistas. Tal organização simplifica o processo de obtenção de dados de uma determinada PI, pois, por meio das revistas, os dados estão agrupados por eventos que ocorreram na semana. Além disso, com o intuito de manter esse conjunto de dados atualizado e relevante nos próximos anos, será apresentado um *workflow* utilizando o GitHub Actions[3], que automatiza o processo de obtenção e processamento das revistas, programado para executar no dia seguinte à publicação da RPI.

Este artigo está organizado como segue. Na Seção 2, são discutidos os trabalhos relacionados. Na Seção 3, é detalhada a metodologia utilizada para a construção do conjunto de dados, incluindo uma descrição do funcionamento do pipeline automatizado. A Seção 4 apresenta uma caracterização do conjunto de dados e, em seguida, a Seção 5 discute sobre possíveis aplicações. Por fim, na Seção 6, o artigo é concluído, apresentando as limitações encontradas e possíveis direções para futuras pesquisas.

2. Trabalhos Relacionados

Até o momento, não foram encontradas referências relacionadas à criação ou à utilização de conjuntos de dados com base nas Revistas de Propriedade Intelectual (RPI) disponibilizadas pelo INPI. No entanto, ao analisar a literatura científica relacionada à elaboração de *datasets* a partir de informações disponibilizadas por entidades públicas, observa-se que os desafios são semelhantes: múltiplas estruturas para os dados, diferentes codificações para arquivos de texto, inconsistência nos valores disponibilizados e dificuldade em manter o conjunto de dados atualizado.

No trabalho de [Barros et al. 2022], foi construído um conjunto de dados utilizando o portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

[2]BuscaWeb: `https://busca.inpi.gov.br/pePI/`

[3]GitHub Actions: `https://github.com/features/actions`

(INEP) para obter informações sobre as instituições de ensino do país . O foco do processamento foi gerar uma série temporal abrangendo o período de 2010 a 2021, visando facilitar a análise dos dados educacionais ao longo do tempo. Como resultado, foram disponibilizadas várias tabelas, totalizando quase três milhões de linhas de dados.

Utilizando os dados disponibilizados pelo Tribunal Superior Eleitoral (TSE), [Vasconcelos et al. 2021] construíram um conjunto de dados sobre candidatos e partidos eleitorais no período de 1945 a 2020. O principal desafio enfrentado foi a falta de padronização dos dados, uma vez que, ao longo do tempo analisado, os dados mais antigos apresentavam erros na codificação dos arquivos texto. Além disso, para enriquecer os dados foram incorporadas informações adicionais, como o gênero dos eleitores, o qual foi inferido por meio da comparação do nome do eleitor com uma base pública de nomes.

Por meio dos dados disponibilizados na plataforma OpenDataSUS, [Gonçalves et al. 2021] construíram diversos conjunto de dados curados relacionados à campanha de vacinação contra COVID-19 no Brasil. O artigo destaca como diferencial a utilização e aplicação dos princípios FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*), os quais asseguram uma melhor qualidade para os dados.

Diferentemente dos estudos mencionados acima, os dados disponibilizados neste estudo reúnem todos os processos de registro de patentes a partir de junho 2018, fornecidos pelo INPI. Outra contribuição deste trabalho é a proposta e implementação de um sistema que automatiza o processo de obtenção de novos dados (i.e., novas revistas), garantindo a atualização contínua do conjunto de dados e a sua relevância para pesquisadores no futuro. Isso permitirá que os pesquisadores tenham acesso a informações atualizadas e precisas sobre os processos de registro de patentes, facilitando suas análises e estudos na área.

3. Metodologia

Um dos principais objetivos deste artigo é automatizar o processo de atualização dos dados. Dessa forma, as etapas para construção do conjunto de dados foram desenhadas para poderem ser executadas semanalmente, de forma eficiente. A Figura 1 ilustra as seguintes etapas: obtenção de revistas faltantes, extração e conversão das revistas, processamento dos dados e atualização do repositório.

Todas as etapas foram implementadas utilizando TypeScript[4] que, devido a sua natureza assíncrona, simplifica o *download* em paralelo das revistas. Além disso, a linguagem oferece funções nativas para manipulação de arquivos JSON[5], o que é essencial para interagir com os arquivos no repositório. Nas próximas seções, serão discutidos os detalhes de implementação de cada etapa.

3.1. Obtenção de revistas faltantes

O objetivo principal desta etapa é verificar se há alguma revista faltando no repositório e, caso seja encontrada alguma, adicioná-la à lista para ser obtida. Dessa forma, pode-se garantir que sempre que esta etapa for executada, todas as revistas a partir da edição 2474 estarão disponíveis localmente no repositório.

[4]TypeScript: `https://www.typescriptlang.org/`
[5]JSON: `https://www.json.org/json-pt.html`

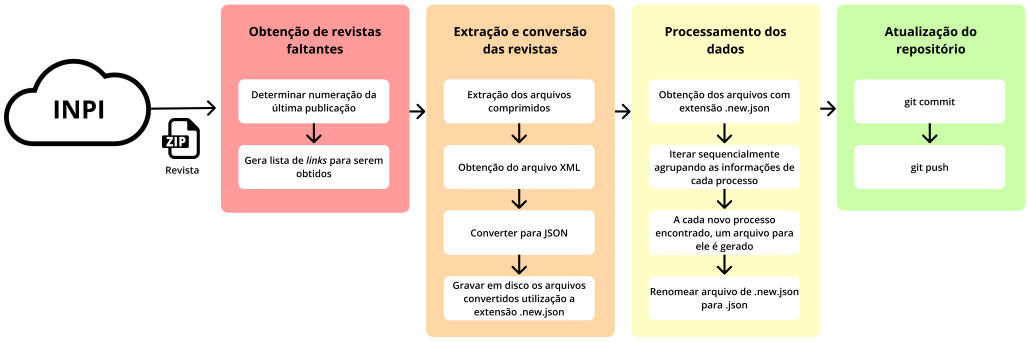

Figura 1. Visualização das etapas para criação do conjunto de dados.

O INPI disponibiliza os arquivos das revistas em um arquivo comprimido. Esses arquivos são mantidos no repositório como *cache* e, por isso, caso seja necessário reprocessar alguma informação, não será necessário obter os arquivos novamente dos servidores do INPI. Além disso, o site no qual as revistas estão disponibilizadas apresenta diversos *links* quebrados para as 1000 primeiras publicações. Dessa forma, o repositório pode servir de alternativa para a obtenção dos arquivos comprimidos caso estes não estejam disponíveis.

Nesta etapa, inicia-se a busca a partir da publicação 2474 por ser a primeira a disponibilizar um arquivo XML dos dados. A partir dela, todas as edições são verificadas sequencialmente até a última publicação possível. Para determinar a numeração da edição mais recente, é utilizado uma RPI de referência e acrescentados em sua edição o número de semanas passadas até o dia atual. Afinal, a cada semana, uma nova revista é publicada.

Além disso, foi necessário definir um limite de acessos para evitar ser bloqueado. Mesmo que o INPI nunca ter se pronunciado sobre impor um limite nas requisições para os servidores provedores de revistas, já foi informado que a plataforma BuscaWeb, administrada pelo INPI, possui um limite de 120 acessos por minuto para *crawlers* [INPI 2023]. Dessa maneira, por segurança, foi estipulado um limite de 100 requisições por minuto para essa etapa.

3.2. Extração e conversão das revistas

Após obter todos os arquivos comprimidos que estavam faltando, é necessário extraí-los. Para isso, foi utilizado a biblioteca ADM-ZIP[6]. O primeiro passo é verificar se no arquivo comprimido existe um arquivo XML. Caso exista, o conteúdo do arquivo é obtido e convertido para JSON utilizando a biblioteca xml2js[7]. O resultado dessa conversão é tratado, removendo algumas estruturas de lista desnecessárias que são comuns de ser incluídas por essa biblioteca. Por fim, o arquivo deve ser persistido em disco para que próximas execuções desta etapa não convertam arquivos compridos já processados.

Além disso, é necessário deixar claro que esse JSON é novo no repositório para que as próximas etapas só processem as informações desses arquivos recém-incluídos. Por isso, todos os arquivos gerados por essa etapa são salvos com a extensão *.new.json*. Por outro lado, caso o arquivo comprimido não possua um arquivo XML, ele é descartado e um aviso é exibido no console.

[6]ADM-ZIP: `https://www.npmjs.com/package/adm-zip`
[7]xml2js: `https://www.npmjs.com/package/xml2js`

3.3. Processamento dos dados

Até o momento, foi realizado somente um pré-processamento dos dados para simplificar a manipulação das revistas. Nesta etapa, os arquivos JSON gerados pela etapa anterior serão processados, visando encontrar informações que ainda não estão disponíveis no conjunto de dados. É importante apontar que neste trabalho, o processamento se refere a uma mudança na estrutura dos dados e não inclui um tratamento dos valores encontrados.

Para isso, gera-se uma lista com os arquivos que possuem a extensão *.new.json*. Cada arquivo é aberto e todos os despachos[8] contidos nele são analisados. Caso o despacho esteja associado a um processo de proteção novo, um arquivo JSON será criado utilizando como nome o número do processo e todas as informações do despacho são incluídas no arquivo.

Por outro lado, caso já exista um processo de pedido com essa numeração no repositório, o arquivo referente a ele é aberto e todas as informações do despacho são incluídas. Se o despacho incluir um dado que já está no arquivo do processo, ele será sobrescrito. Por isso, é fundamental utilizar a ordem de publicação das revistas para o processamento, de forma que o dado mais recente seja o último a ser adicionado.

Todos os despachos identificados também são incluídos no arquivo. O campo *despachos* apresenta uma lista com o código, título e revista do despacho. Dessa forma, é possível saber exatamente quais revistas referenciam aquele processo, sem precisar iterar por todas as publicações.

Vale ressaltar que esse processamento deve ser feito sequencialmente para garantir a integridade das informações. Caso esse processamento seja feito em paralelo, seria difícil incluir novas informações nos arquivos já que múltiplas *threads* estariam com ele aberto em diferentes estados.

Por fim, os arquivos JSON processados são renomeados para trocar a extensão de *.new.json* para *.json*. Dessa forma, os arquivos podem ficar salvos no repositório e nunca serem reprocessados sem necessidade.

3.4. Atualização do repositório

Após obter as revistas faltantes e modificar arquivos no repositório, é necessário persistir esses dados para as alterações ficarem disponíveis.

Como os dados estão em um repositório Git, basta executar os comando *git commit* seguido de um *git push*. Com isso, todas as alterações serão persistidas e enviadas para o GitHub para serem disponibilizadas publicamente.

3.5. *Workflow* no GitHub Actions

O GitHub *Actions* ajuda a automatizar os fluxos de trabalho de desenvolvimento de software dentro do GitHub. Em outras palavras, com essa ferramenta é possível definir uma sequência de passos que devem ser executados em uma máquina virtual remota, de forma gratuita (em repositórios públicos), visando atingir um determinado objetivo.

[8]Um despacho é um evento/exigência que ocorre em um processo de pedido da Propriedade Intelectual que o INPI informa na revista

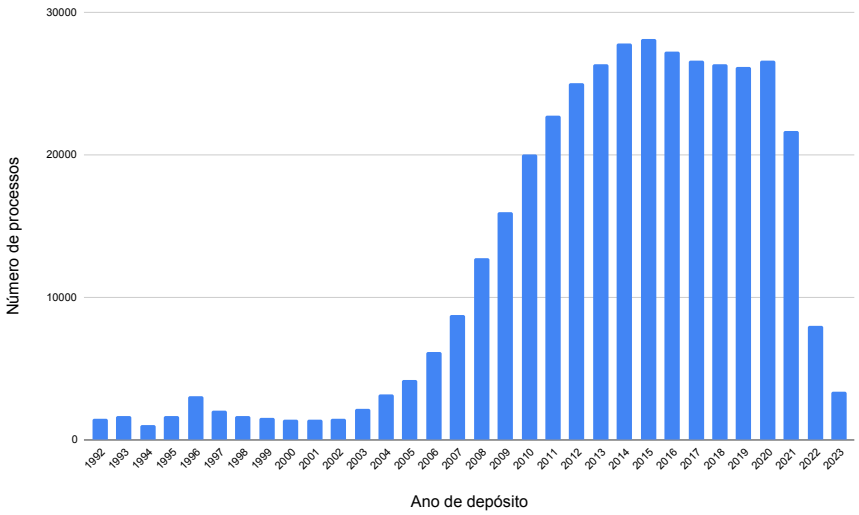

Figura 2. Número de processos depositados por ano.

Devido a essa funcionalidade, foi possível criar um fluxo que executa as etapas descritas todas as quartas-feiras de manhã, o que garante que os dados se mantenham atualizados sem necessidade de intervenção manual.

4. Caracterização do Conjunto de Dados

Para a construção do conjunto de dados, foi necessário baixar e processar 264 revistas e analisar 1.507.796 despachos. Dessa maneira, foi possível identificar e agregar informações de 408.705 processos de proteção de PI. É importante ressaltar que, devido ao fluxo automatizado, comentado na Seção 3.5, o número de processos neste conjunto de dados muda frequentemente, já que uma nova revista é incluída semanalmente. A Tabela 1 inclui dados sobre as informações coletadas. Vale ressaltar, que dos despachos analisados, dois não tinham um número de processo e foram ignorados no processamento.

Tabela 1. Informações relevantes sobre o conjunto de dados desenvolvido.

Número de Processos	408.705
Número de Titulares	113.628
Número de Titulares Brasileiros	47.390
Número de Inventores	411.724

Uma outra maneira de descrever o conjunto de dados é analisando as datas de depósito de todos os processos. Essa data indica o momento em que o processo foi inserido no sistema do INPI. A Figura 2 apresenta um gráfico em barras com o número de processos em relação o ano de depósito. Nela, é possível perceber que o conjunto de dados apresenta, em sua maioria, processos depositados entre 2007 e 2021, tendo um valor máximo de 28.134 registros em 2015. É importante ressaltar que aproximadamente 5% dos processos não possuíam uma data de depósito em nenhum despacho publicado e foram desconsiderados na Figura 2.

Tabela 2. Dicionário de dados para o conjunto de dados deste trabalho.

Campo	Tipo	Descrição
numero	string	Número do processo
dataDeposito	string	Data em que ocorreu o depósito do processo.
dataConcessao	string	Data em que ocorreu a concessão da patente.
dataFaseNacional	string	Data que ocorrerá a fase nacional
pedidoInternacional.numeroPCT	string	Número do pedido internacional
pedidoInternacional.dataPCT	string	Data de depósito do pedido internacional
publicacaoInternacional.numeroOMPI	string	Número do processo no sistema da OMPI
publicacaoInternacional.dataOMPI	string	Data de depósito do pedido no sistema da OMPI
titulo	string	Título da patente
IPC	string[]	Lista com a classificação atribuídas seguindo as categorias do IPC
titulares	object[]	Lista com informações sobre os titulares das patentes
titulares.nomeCompleto	string	Nome completo do titular
titulares.pais	string	Sigla para o país de origem do titular
titulares.uf	string	Caso o titular seja brasileiro, este campo indica qual é o estado de origem do titular
inventores	string[]	Lista com o nome dos inventores da patente
prioridadesUnionistas	object[]	Lista com informações sobre a prioridade da patente
prioridadesUnionistas.siglaPais	string	Sigla do país de origem da prioridade
prioridadesUnionistas.numeroPrioridade	string	Número da prioridade
prioridadesUnionistas.dataPrioridade	string	Data da prioridade unionista
divisaoPedido.dataDeposito	string	Data de deposito da divisão de pedido
divisaoPedido.numero	string	Numero do processo que foi dividido
pedidoPrincipal.dataDeposito	string	Data de depósito do pedido de complemento de um processo já existente.
pedidoPrincipal.numero	string	Número do processo que teve alterações
despachos	object[]	Lista com os despachos que referenciam o processo
despachos.codigo	string	Código do despacho
despachos.titulo	string	Descrição sobre o tipo do despacho
despachos.rpi	number	Número da revista em que o despacho foi encontrado

Além de incluir as informações sobre os processos, o repositório também apresenta um conjunto de dados auxiliar que inclui a descrição para cada classificação de patente. Os dados foram extraídos do site oficial da World Intellectual Property Organization (WIPO)[9] e contam com 78.378 categorias.

4.1. Dicionário de dados

O INPI não apresenta uma documentação adequada que descreva o significado de cada campo encontrado nos despachos das Revistas de Propriedade Industrial. Isso dificulta a compreensão do conjunto de dados e desmotiva aqueles que não possuem um conhecimento aprofundado no processo de registro de patentes.

Visando contornar esse problema, foi incluído no repositório um dicionário de dados que descreve o significado de cada campo existente nos arquivos agregados de processos. A Tabela 2 demonstra o conteúdo deste dicionário.

5. Aplicações

Diversas aplicações podem ser realizadas utilizando o conjunto de dados apresentado. As Seções 5.1 e 5.2 descrevem alguns exemplos.

5.1. Influência de empresas estrangeiras no registro de patentes de 2012 a 2020

Por meio do campo *titulares.pais*, foi possível identificar se há um aumento da influência de empresas estrangeiras no registro de patentes no Brasil. A Figura 3 demonstra o número de processos em que os titulares são brasileiros por ano em comparação com os de titulares estrangeiros.

[9]WIPO: `https://www.wipo.int/portal/en/index.html`

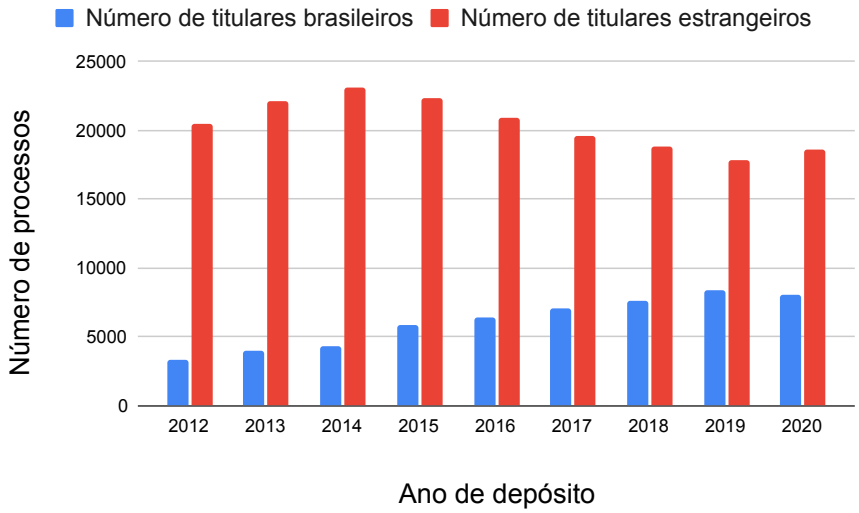

Figura 3. Número de processos que possuem titulares brasileiros e estrangeiros por ano de depósito.

Nota-se que apesar do número de titulares brasileiros ter aumentado ao longo dos anos, os titulares estrangeiros são a maioria dos registros em todos os anos.

5.2. Área na qual houve o maior número de registro de patentes de 2012 a 2020

Utilizando o campo *IPC*, é possível identificar qual a categoria do processo dentro da Classificação Internacional de Patentes (IPC). Por conta de ser um sistema hierárquico, pode-se reduzir as classificações do IPC em oito grandes áreas: necessidades humanas, operações de processamento e transporte, química e metalurgia, têxteis e papel, construções fixas, engenharia mecânica, física e eletricidade. A Figura 4 demonstra o número de patentes por área. Nota-se que a categoria de Necessidades Humanas foi a classificação com maior número de registros de patentes no período.

6. Conclusões

Este artigo apresentou um conjunto de dados inédito para a comunidade científica com informações sobre os registros de patentes desde junho de 2018. Os dados encontrados nas Revistas de Propriedade Industrial foram processados e separados em arquivos individuais para cada processo, agregando todas as informações encontradas. Vale destacar que o *workflow* apresentado utilizando o GitHub *Actions* irá garantir a atualização contínua dos dados do repositório, mantendo-os sempre atualizado.

Uma limitação do conjunto de dados é que ele apresenta apenas informações de revistas a partir de junho de 2018. Antes dessa data, os dados eram disponibilizados em arquivos de texto sem uma estrutura bem definida ou documentada. Portanto, o processamento dessas revistas exigiria o desenvolvimento de um *parser* específico para esses arquivos de texto.

Como trabalhos futuros, planeja-se incluir informações sobre outros tipos de propriedades intelectuais. Além de informações sobre registro de patentes, pretende-se incluir dados sobre registros de programas de computador.

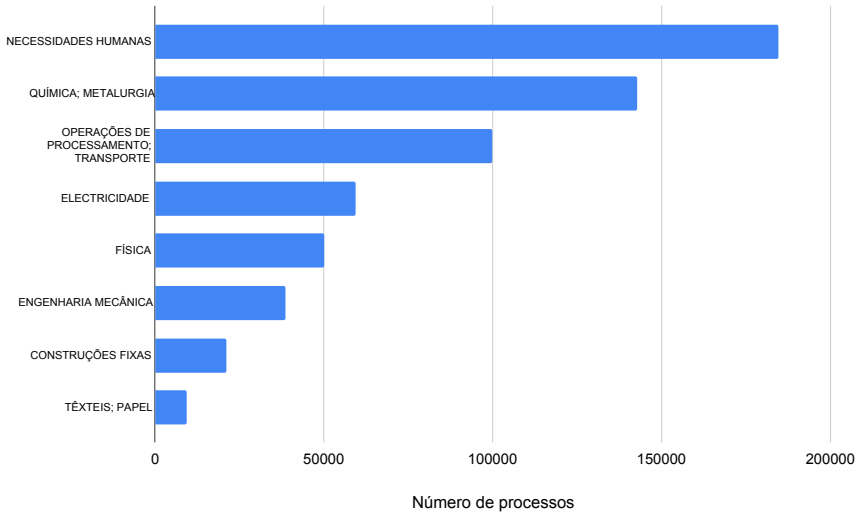

Figura 4. Número de processos por categoria.

O conjunto de dados está disponível em `https://github.com/cie-cefet-mg/inpi-db`. Nesse repositório do GitHub, é possível visualizar as execuções do *workflow* na aba *Actions*, e os dados podem ser encontrados na pasta *data* na raíz do projeto.

Referências

Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. *Anais do IV Dataset Showcase Workshop (DSW 2022)*, pages 35–45.

Gonçalves, M. V. F., dos Santos, J. S., Ferreira, C. Z., Zavaleta, J., da Cruz, S. M. S., and Sampaio, J. O. (2021). Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. *Anais do III Dataset Showcase Workshop (DSW 2021)*, pages 148–159.

INPI (2020). Guia básico — instituto nacional da propriedade industrial. Disponível em: `https://www.gov.br/inpi/pt-br/servicos/patentes/guia-basico`. Acesso em: 29 de junho de 2023.

INPI (2023). Inpi bloqueia acessos de robôs aos sistemas de forma automatizada. Disponível em: `https://www.gov.br/inpi/pt-br/central-de-conteudo/noticias/inpi-bloqueia-acessos-de-robos-aos-sistemas`. Acesso em: 28 de junho de 2023.

Vasconcelos, F. F., Tavares, J. V. S., Ribeiro, M. U., Coutinho, F. J., and Clarindo, J. P. (2021). Candidata: um dataset para análise das eleições no brasil. *Anais do III Dataset Showcase Workshop (DSW 2021)*, pages 160–168.