

Procedimento metodológico para coleta de dados e elaboração de painel de Grupos de Pesquisa CNPq

Data de Início do projeto: 1975 a junho de 2025

1. Tema e Contexto

O Observatório Fiocruz em Ciência, Tecnologia e Inovação em Saúde é uma iniciativa da Coordenação de Informação e Comunicação da Vice-Presidência de Educação, Informação e Comunicação da Fundação Oswaldo Cruz (Fiocruz). Sua proposta é transformar dados provenientes de diversas fontes em informações com alto valor agregado para a Instituição e para a sociedade através da construção de indicadores que tenham como premissa favorecer a tomada de decisão e proporcionar maior transparência à sociedade, fortalecendo assim direta e indiretamente as ações que compõe o Sistema Único de Saúde. Este documento descreve os procedimentos adotados na coleta, transformação e disponibilização do indicador de Grupos de Pesquisa CNPq. Dashboards ou painéis são formatos de visualizações que permitem explorar os dados e analisar as relações entre eles a partir de simples cliques e/ou filtros interativos. O objetivo desta descrição metodológica, além de dar transparência ao indicador é promover o acesso à informação e a reprodutibilidade da coleta e do processamento do conjunto de publicações que compõe Grupos de Pesquisa CNPq no período de cobertura de dados informado.

2. Fontes de Dados

Esta seção descreve a plataforma de origem, o conteúdo recuperado e a estratégia de acesso utilizada na construção do indicador.

2.1. Plataforma/Sistema

Os dados utilizados são públicos e estão disponíveis no Diretório dos Grupos de Pesquisa no Brasil, mantido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), na Plataforma Lattes. O diretório reúne informações cadastrais e acadêmicas sobre grupos de pesquisa vinculados a instituições brasileiras.

2.2. Conteúdo Agregado

A coleta contemplou os seguintes campos: identificador do grupo de pesquisa, nome do grupo, situação do grupo, ano de formação, líder ou líderes, instituição do grupo, unidade federativa, linhas de pesquisa, colaboradores estrangeiros, área de pesquisa, subárea de pesquisa, nomes dos pesquisadores e titulação dos pesquisadores.

Não foram aplicados filtros de data. Foram considerados todos os grupos de pesquisa ativos e inativos registrados desde 1975 até a data da coleta, realizada em 17 de junho de 2025.

2.3. Acesso

Como não foi identificado serviço web para extração direta dos dados do Diretório dos Grupos de Pesquisa, foi utilizado um programa web crawler desenvolvido especificamente para percorrer e indexar o conteúdo do site. O programa recuperou os dados do diretório independentemente do vínculo institucional, armazenando inicialmente o conteúdo em formato HTML, posteriormente convertido para Excel.

3. Procedimento Metodológico para Coleta e Tratamento

Esta seção detalha as etapas de extração, conversão, filtragem, padronização e consolidação dos dados empregados no painel.

3.1. Coleta dos Dados

A coleta foi executada em 17 de junho de 2025 e organizada em duas etapas principais, cada uma responsável por uma parte distinta do fluxo de obtenção e preparação dos registros.

- **Etapla 1 — extração do diretório:** o web crawler percorreu o Diretório dos Grupos de Pesquisa do CNPq e capturou os registros disponíveis, sem restringir a coleta por instituição. Os conteúdos recuperados em HTML foram convertidos para uma planilha Excel.
- **Etapla 2 — seleção institucional:** um segundo programa leu a planilha resultante e aplicou um filtro ao campo “Instituição do grupo”, buscando o termo “Fundação Oswaldo Cruz – Fiocruz” para selecionar os grupos vinculados à instituição.
- **Recorte temporal:** não houve filtro por ano de formação ou situação. Permaneceram na base os grupos ativos e inativos registrados desde 1975 até a data da coleta.

3.2. Tratamento dos Dados

Após a seleção institucional, os registros foram tratados para padronizar as informações de unidade e situação dos grupos, permitindo a consolidação das categorias utilizadas na visualização.

3.2.1. Filtragem Inicial

A filtragem inicial restringiu a base aos grupos cujo campo “Instituição do grupo” continha o termo “Fundação Oswaldo Cruz – Fiocruz”. Dessa forma, foram excluídos da base final os registros pertencentes a outras instituições recuperados pelo web crawler.

3.2.2. Padronização e Consolidação

Os nomes das unidades da Fiocruz associadas aos grupos selecionados foram padronizados. Quando a unidade não havia sido informada pelo líder do grupo no cadastro do CNPq, foi mantida a identificação institucional mais geral, “Fiocruz”.

As situações originais dos grupos foram consolidadas conforme as seguintes regras:

- **Ativo:** inclui os registros classificados no diretório como “Aguardando certificação”, “Certificado” e “Certificado - Não-atualizado há mais de 12 meses”.

- **Em preenchimento:** mantém os registros originalmente classificados como “Em preenchimento”.
- **Inativo:** inclui os registros originalmente classificados como “Excluído”.

3.2.3. Armazenamento

Os conteúdos extraídos foram inicialmente armazenados em HTML e posteriormente convertidos para uma planilha Excel. Após a filtragem e a padronização, os resultados foram consolidados nessa estrutura tabular para utilização como base do painel.

4. Elaboração do Painel/Relatório

Após a consolidação dos resultados, a planilha foi utilizada como fonte para a construção de uma visualização interativa publicada no Portal do Observatório Fiocruz em Ciência, Tecnologia e Inovação em Saúde.

4.1. Princípios e Fontes

A elaboração do painel utilizou como fonte a planilha Excel resultante da extração, da filtragem institucional e da padronização dos dados. A estrutura do indicador foi organizada para permitir a exploração das características dos grupos de pesquisa vinculados à Fiocruz e das pessoas e áreas associadas a esses grupos.

4.2. Identidade Visual e Funcionalidade

A visualização interativa foi publicada no site do Observatório Fiocruz. A publicação tem como finalidade facilitar a consulta e a análise dos dados dos grupos de pesquisa da Fiocruz.

5. Análise Específica: Grupos de Pesquisa da Fiocruz

A base consolidada possibilita analisar a distribuição dos grupos de pesquisa por situação, ano de formação, unidade da Fiocruz, unidade federativa, área e subárea de pesquisa. Também permite examinar as linhas de pesquisa, as lideranças, os pesquisadores, suas titulações e a presença de colaboradores estrangeiros.

A padronização das unidades e o agrupamento dos status favorecem comparações institucionais e temporais. Ao mesmo tempo, a manutenção da categoria geral “Fiocruz” para registros sem unidade declarada preserva os dados recuperados sem atribuir uma vinculação não informada no cadastro original.

6. Considerações Técnicas

Esta seção reúne os desafios técnicos, as estratégias adotadas e as limitações documentais do procedimento.

- **Ausência de serviço de extração direta:** a utilização de web crawler foi necessária porque não foi identificado serviço web para recuperação estruturada dos dados do diretório.
- **Abrangência da coleta:** o crawler recuperou dados de grupos independentemente do vínculo institucional; a seleção dos registros da Fiocruz ocorreu em etapa posterior, por meio de filtro aplicado à instituição do grupo.
- **Conversão de formato:** os registros capturados em HTML foram convertidos para Excel para viabilizar a leitura pelo segundo programa, o tratamento e a integração ao painel.
- **Classificação das situações:** os status originais foram agrupados em Ativo, Em preenchimento e Inativo de acordo com regras explicitadas nesta metodologia.
- **Data da coleta:** os resultados representam o conteúdo disponível no diretório em 17 de junho de 2025 e podem sofrer alterações em coletas posteriores, conforme a atualização dos cadastros pelo CNPq e pelos responsáveis pelos grupos.