

## **Procedimento metodológico para coleta de dados e elaboração de painel dos Programas de Pós-Graduação Fiocruz**

*Referência: Painel-Educação*

*Data de Início do projeto: 2017 a dezembro de 2024*

### **1. Tema e Contexto**

O Observatório em Ciência, Tecnologia e Inovação em Saúde (Observatório CT&I 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 Programas de Pós-Graduação Fiocruz. 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 Programas de Pós-Graduação Fiocruz no período de cobertura de dados informado.

## **2. Fontes de Dados**

Esta seção descreve o sistema de origem dos dados, o conteúdo disponibilizado e a estratégia utilizada para realizar a consulta sistemática das informações.

### **2.1. Plataforma/Sistema**

A Plataforma Sucupira é mantida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e constitui-se como a ferramenta oficial de acompanhamento dos Programas de Pós-Graduação brasileiros. Para o indicador, foram recuperados os registros dos Programas de Pós-Graduação vinculados à Fundação Oswaldo Cruz.

### **2.2. Conteúdo Agregado**

A plataforma agrega informações sobre a estrutura dos programas, dados básicos e situação dos cursos, áreas de concentração, linhas de pesquisa, instituições e cursos vinculados. Também reúne dados sobre docentes, discentes, egressos, pós-doutorandos, participantes externos, teses e dissertações, produções científicas e projetos de pesquisa concluídos ou em andamento.

### **2.3. Acesso**

Embora os dados sejam públicos, sua consulta sistemática exige navegação por múltiplas abas, filtros, tipos de vínculo, graus acadêmicos, anos-base e subtipos de produção. Por esse motivo, foi desenvolvido um conjunto de programas em Python com automação baseada em Selenium. O Selenium controla o navegador e permite ao código executar ações como abrir páginas, preencher campos, selecionar filtros, clicar em links e recuperar os conteúdos exibidos.

### 3. Procedimento Metodológico para Coleta e Tratamento

Esta seção detalha o processo de extração, filtragem, deduplicação, organização e armazenamento dos dados utilizados na elaboração do painel.

#### 3.1. Coleta dos Dados

Para recuperar as informações dos Programas de Pós-Graduação da Fiocruz, a coleta foi dividida em cinco etapas, executadas por módulos específicos. A entrada comum a todos os programas é um arquivo tabular com os campos ppg\_id e path\_id, identificadores únicos atribuídos pela CAPES aos programas no momento de seu registro na Plataforma Sucupira.

##### 3.1.1. Etapa 1 — Situação Atual e Dados Básicos

O primeiro módulo realiza a navegação para cada ppg\_id e recupera os dados gerais dos programas. A execução utiliza espera para o carregamento das páginas, repetição de tentativas em caso de falha e armazenamento dos resultados em arquivos .csv separados.

- **Dados básicos do programa:** nome, nota, coordenador, áreas de conhecimento, modalidade e situação do curso.
- **Estrutura acadêmica:** áreas de concentração, linhas de pesquisa, instituições envolvidas e lista de cursos vinculados ao Programa de Pós-Graduação.

##### 3.1.2. Etapa 2 — Pessoas Vinculadas

O segundo módulo recupera os perfis das pessoas associadas a cada programa, segmentando os registros por tipo de vínculo e grau acadêmico.

- **Públicos recuperados:** docentes, discentes, egressos, pós-doutorandos e participantes externos.
- **Campos coletados:** nome, nacionalidade, país de titulação, instituição de vínculo e ano de referência da coleta.

##### 3.1.3. Etapa 3 — Teses e Dissertações

O terceiro módulo coleta, para cada programa e ano-base, títulos, autores e identificadores únicos das teses e dissertações, além do tipo, subtipo, programa e instituição de origem. Esses registros permitem examinar padrões de formação, orientações e temas recorrentes de pesquisa.

##### 3.1.4. Etapa 4 — Produções Científicas

O quarto módulo realiza varreduras por subtipo de produção, incluindo artigos completos, livros, patentes e outras categorias disponibilizadas pela plataforma. São extraídos os

identificadores das produções, títulos, autores, programas, instituições, tipos, subtipos e anos de referência.

### **3.1.5. Etapa 5 — Projetos de Pesquisa**

O quinto módulo explora os projetos registrados por cada Programa de Pós-Graduação, considerando filtros de natureza, como individual ou em rede, e de situação, como em andamento ou concluído.

- **Campos coletados:** título, área de concentração, linha de pesquisa, responsável pelo projeto, programa, instituição, natureza e situação do projeto.

## **3.2. Tratamento dos Dados**

Como uma mesma produção ou um mesmo projeto pode aparecer em diferentes Programas de Pós-Graduação em razão de colaborações intra-PPG e inter-PPGs, foi adotado um procedimento automatizado para identificar e consolidar registros redundantes. O tratamento segue padrões empregados em outros projetos do Observatório da Fiocruz.

### **3.2.1. Filtragem Inicial**

Antes da deduplicação, a base foi restringida aos registros efetivamente relacionados à Fiocruz. Após a coleta, as produções científicas e os projetos foram verificados para manter apenas aqueles em que pelo menos um coautor estivesse vinculado a um Programa de Pós-Graduação da instituição.

A verificação compara os nomes dos coautores de cada registro com uma lista institucional interna que reúne pessoas associadas à Fiocruz, incluindo docentes, colaboradores, discentes, egressos e outros vínculos. Esse procedimento busca assegurar que a base analítica represente a contribuição científica institucional.

### **3.2.2. Deduplicação**

A deduplicação identifica registros que aparecem em diferentes contextos, mas representam a mesma produção ou o mesmo projeto. Foi utilizada uma medida de similaridade textual baseada na distância de Levenshtein, considerando equivalentes os títulos com pelo menos 93% de similaridade.

Antes da comparação, os títulos foram normalizados mediante conversão para letras minúsculas e remoção de acentos, pontuações e espaços. A comparação foi restrita ao mesmo tipo de registro e ao mesmo ano, reduzindo o risco de associações indevidas entre categorias distintas.

### **3.2.2.1. Produções Científicas**

Quando duas ou mais ocorrências de uma produção foram identificadas como duplicadas, elas foram agrupadas em um único registro. Os identificadores ppg\_id e producao\_id foram reunidos, ordenados e apresentados separados por ponto e vírgula. Como título final, foi mantida a versão mais completa, definida pelo maior comprimento de caracteres. Assim, cada produção é contabilizada apenas uma vez, preservando os identificadores associados.

### **3.2.2.2. Projetos de Pesquisa**

Para os projetos, foi aplicado o mesmo princípio de deduplicação. Os registros equivalentes foram consolidados em uma única entrada, reunindo campos como ppg\_id, situação, responsáveis e códigos do projeto, apresentados em ordem crescente e separados por ponto e vírgula. O título mais completo foi mantido como referência final, evitando superestimação por repetição de registros.

### **3.2.3. Armazenamento**

Os resultados das coletas foram exportados em arquivos .csv com separador TAB e codificação UTF-8. Os arquivos empregados no painel foram armazenados em uma pasta compartilhada no SharePoint da Fiocruz, permitindo integração dinâmica com o Power BI e maior segurança no armazenamento.

Entre as fontes utilizadas estão arquivos de cursos, pessoas vinculadas, projetos, produções científicas, teses, dados institucionais dos programas, o arquivo fiocruz-ppgs2025.csv e a planilha complementar “Lista PPG-SS Fiocruz\_2025.xlsx”.

#### 4. Elaboração do Painel/Relatório

Com o objetivo de ampliar a compreensão dos dados extraídos da Plataforma Sucupira e facilitar a análise exploratória das informações dos Programas de Pós-Graduação da Fiocruz, foi desenvolvido um painel interativo no Power BI Desktop, versão 2.143.878.0 (64-bit), de maio de 2025.

##### 4.1. Princípios e Fontes

A elaboração do painel seguiu princípios de acessibilidade, padronização visual e clareza informacional. Foram priorizados filtros interativos e gráficos de leitura simples, inclusive para usuários com pouca familiaridade com ferramentas computacionais.

As visualizações foram construídas de forma colaborativa pela equipe do Observatório em Ciência, Tecnologia e Inovação em Saúde e por representantes da área de Educação da VPEIC/Fiocruz, buscando coerência entre a apresentação dos indicadores e os objetivos institucionais.

##### 4.2. Modelagem e Transformação

As transformações foram conduzidas majoritariamente no Power Query, com intervenções voltadas à padronização de campos textuais e à estruturação relacional. Os maiores tratamentos foram aplicados ao arquivo fiocruz-ppgs2025.csv e à planilha “Lista PPG-SS Fiocruz\_2025.xlsx”.

- **Transformações principais:** remoção de espaços e linhas vazias, promoção de cabeçalhos, preenchimento de campos ausentes e separação de colunas compostas.

A modelagem considerou o arquivo fiocruz-ppgs2025.csv como tabela fato, relacionado às demais tabelas dimensão por meio do campo ppg\_id, também referenciado como Cód CAPES. Os relacionamentos foram configurados conforme a cardinalidade das entidades, com relações um-para-muitos e um-para-um quando aplicável.

##### 4.3. Identidade Visual e Funcionalidade

A identidade visual foi construída segundo o Manual da Marca da Fiocruz, versão de agosto de 2024, que orienta o uso de cores institucionais, fontes tipográficas e princípios de legibilidade. A aplicação dessas diretrizes reforça a identidade da Fundação e a padronização entre seus produtos de informação.

O painel contempla uma aba exclusiva de filtros, permitindo selecionar variáveis como ano de produção, área básica, vínculo, unidade e grau acadêmico. A navegação foi organizada

para apoiar a descoberta de informações, fomentar a cultura de dados e subsidiar análises estratégicas e tomadas de decisão baseadas em evidências no âmbito da pós-graduação da Fiocruz.

## 5. Análises Avançadas — Redes de Colaboração

As redes de colaboração permitem compreender a dinâmica de interação intra-PPG e inter-PPGs. A partir das produções científicas e dos projetos previamente filtrados e deduplicados, é possível estruturar redes em que cada vértice corresponde a um Programa de Pós-Graduação e cada aresta representa a existência de produção ou projeto compartilhado.

Essa representação pode apoiar a identificação de programas mais conectados, núcleos de cooperação, vínculos recorrentes e oportunidades de integração institucional. A qualidade da rede depende da correta consolidação dos identificadores e da preservação das associações entre os registros e os respectivos programas.

## 6. Considerações Técnicas

A coleta automatizada envolve desafios relacionados à estabilidade da Plataforma Sucupira, ao tempo de carregamento das páginas e ao volume de combinações de filtros. Para aumentar a robustez do processo, foram adotados os seguintes cuidados:

- **Cuidados na automação:** uso de tempos de espera aleatórios entre requisições para reduzir o risco de bloqueios por uso excessivo e permitir o carregamento completo das páginas.
- **Tolerância a falhas:** repetição automática de tentativas, limitada a cinco execuções por falha, em casos de timeout ou problemas de carregamento.
- **Estratégia de execução:** divisão da coleta em etapas e parametrização do ano-base, do tipo de vínculo, do grau acadêmico e dos subtipos de filtros, permitindo execução distribuída e incremental.
- **Verificação de completude:** recomendação de múltiplas execuções, pois instabilidades temporárias da plataforma podem ocasionar perdas momentâneas de registros.
- **Padronização da saída:** exportação em arquivos .csv com separador TAB e codificação UTF-8, facilitando a integração com Power BI, R, Python, Excel e ferramentas de bibliometria.
- **Integração institucional:** armazenamento dos arquivos no SharePoint da Fiocruz e conexão com o Power BI para centralizar as fontes utilizadas pelo painel.