

Procedimento metodológico para coleta de dados e elaboração de painel de “Produção científica com parceiros internacionais”

Data de Início do projeto: Janeiro de 2008 a dezembro de 2025

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 produção científica com parceiros internacionais da Fiocruz em um dashboard. 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 produção científica com parceiros internacionais no período de cobertura de dados informado.

1.1. Tema

A análise da produção científica da Fundação Oswaldo Cruz (Fiocruz) com participação de instituições estrangeiras requer informações confiáveis e atualizadas sobre publicações, autores, afiliações, parceiros e países. A

sistematização desses dados permite acompanhar a inserção internacional da produção institucional e subsidiar análises estratégicas em pesquisa, ciência, tecnologia, inovação e relações internacionais.

1.2. Contexto

Nesse contexto, o Observatório em Ciência, Tecnologia e Inovação em Saúde (Observatório CT&I em Saúde), iniciativa da Coordenação de Informação e Comunicação da Vice-Presidência de Educação, Informação e Comunicação da Fiocruz, transforma dados de diferentes fontes em informações de alto valor agregado. Seus indicadores e painéis interativos contribuem para a tomada de decisão, a transparência institucional e o fortalecimento das ações relacionadas ao Sistema Único de Saúde.

1.3. Objetivo do Trabalho e Escopo

O trabalho tem como objetivo identificar, organizar e disponibilizar as produções científicas da Fiocruz que possuem ao menos um parceiro internacional. O indicador deriva do painel de Produção Científica e utiliza os mesmos procedimentos de coleta, tratamento, harmonização, enriquecimento, armazenamento e curadoria, acrescidos de um filtro específico para colaboração internacional. O escopo abrange janeiro de 2008 a junho de 2025 e busca ampliar o acesso à informação, a reprodutibilidade do processo e as análises sobre a cooperação científica da instituição.

2. Fontes de Dados

Esta seção descreve os sistemas, plataformas e repositórios utilizados na formação do conjunto de dados e o tipo de informação que eles contêm.

2.1. Plataformas e Sistemas

A base do indicador é constituída por registros de publicações científicas coletados em sete fontes: Web of Science (WoS), Scopus, PubMed, Literatura Latino-Americana e do Caribe em Ciências da Saúde (Lilacs), Scientific Electronic Library Online (SciELO), Repositório Institucional Arca e Plataforma Lattes. Também são utilizadas fontes complementares, como ORCID, OpenAlex, Directory of Open Access Journals (DOAJ), Crossref e Research Organization Registry (ROR), para enriquecer, validar e harmonizar os registros.

2.2. Conteúdo Agregado

As fontes agregam informações sobre títulos, anos de publicação, tipologias documentais, autores, afiliações institucionais, unidades da Fiocruz, instituições parceiras, países, veículos de publicação, palavras-chave e identificadores como DOI, ISSN, ORCID e ResearcherID. Quando disponíveis, também são incorporadas informações sobre áreas do conhecimento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), acesso aberto e metadados de instituições e periódicos.

2.3. Acesso

A coleta é realizada de acordo com as características de cada fonte. WoS, Scopus, PubMed e Lilacs são consultadas mediante estratégias de busca por afiliação institucional adaptadas às respectivas interfaces. A SciELO é processada a partir de registros em XML; os dados do Arca são disponibilizados em formato tabular por parceria interna com o Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT/Fiocruz); e os currículos Lattes de servidores ativos e inativos são consultados por meio da Base Lattes Fiocruz, a partir de dados fornecidos pela área de gestão de pessoas. A consulta sistemática exige procedimentos automatizados e de harmonização devido ao volume de registros, à

diversidade de formatos e às múltiplas formas de declaração das afiliações institucionais.

3. Procedimento Metodológico para Coleta e Tratamento

Esta seção detalha o processo de extração, limpeza, organização, harmonização e seleção dos dados utilizados no indicador.

3.1. Coleta dos Dados

A produção científica com parceiros internacionais não é coletada em um processo independente. O indicador é derivado do universo de publicações que compõe o painel de Produção Científica da Fiocruz. A coleta e a preparação desse universo podem ser organizadas nas seguintes etapas:

- **Busca e recuperação:** aplicação de uma string de busca por afiliação, adaptada às interfaces das bases bibliográficas, para localizar publicações nas quais a Fiocruz ou suas unidades estejam declaradas entre as afiliações dos autores.
- **Importação de fontes institucionais:** incorporação de registros do Arca e dos currículos Lattes dos servidores da Fiocruz, observando o período de vínculo institucional quando essa informação está disponível.
- **Conversão de formatos:** transformação dos arquivos de origem - como CSV, RIS, PubMed Data, XML e planilhas - para uma estrutura intermediária padronizada em JSON.
- **Harmonização e enriquecimento:** aplicação de dicionários, algoritmos e fontes complementares para padronizar autores, afiliações, parceiros, países, veículos, tipologias e identificadores.
- **Seleção do indicador:** aplicação do critério de colaboração internacional sobre o conjunto tratado, mantendo as publicações que possuem ao menos um parceiro internacional identificado.

3.2. Tratamento dos Dados

Para garantir a representação de uma publicação única e reduzir inconsistências entre registros provenientes de diferentes fontes, foi adotado um fluxo de tratamento voltado à validação, padronização, harmonização e consolidação das informações. Os procedimentos incluem correção de caracteres e codificação,

remoção de tags e pontuações indevidas, padronização de espaços, normalização de nomes e aplicação de dicionários de referência.

3.2.1. Filtragem Inicial

Antes da consolidação, são removidos registros sem título válido, sem ano de publicação ou sem autoria, bem como registros fora do recorte temporal do indicador. Também são desconsideradas tipologias que não representam produção científica inédita, como erratas, correções, comentários, respostas de autores e retratações. Após a harmonização das afiliações, permanecem apenas as publicações com ao menos um autor vinculado à Fiocruz. Para a composição do painel de parceiros internacionais, aplica-se ainda o filtro que mantém somente as produções classificadas como possuindo ao menos uma instituição parceira internacional.

3.2.2. Deduplicação

A deduplicação busca impedir que uma mesma publicação, recuperada em mais de uma fonte, seja contabilizada repetidamente. Cada registro recebe um identificador persistente, denominado `fiocruz_id`, gerado a partir do título e do ano de publicação. Na inserção no banco de dados, a existência prévia da publicação é verificada por meio de critérios como identidade do `fiocruz_id`, DOI, título e ano, ou título e veículo de publicação, com normalização de letras minúsculas e acentuação quando necessário.

- **Registros inéditos:** quando a publicação ainda não existe no banco, o registro é inserido e relacionado aos autores, afiliações, parceiros e demais dimensões correspondentes.
- **Registros já existentes:** quando a publicação já foi identificada, não é criada uma nova ocorrência; a informação sobre a base de origem é atualizada para preservar a rastreabilidade da coleta.

3.2.3. Armazenamento

Os dados tratados são armazenados em banco de dados relacional PostgreSQL. A carga é realizada por scripts desenvolvidos em Python, com mapeamento objeto-

relacional por meio da biblioteca SQLAlchemy. A ordem de precedência das fontes considera a qualidade e a completude dos registros, priorizando WoS, Scopus, PubMed, SciELO, Lilacs, Arca e, por último, Plataforma Lattes. Os arquivos JSON harmonizados funcionam como estrutura intermediária entre a coleta e a carga no banco de dados.

4. Elaboração do Painel/Relatório

Com o objetivo de ampliar a compreensão sobre as colaborações científicas internacionais da Fiocruz e facilitar a análise exploratória, foi desenvolvido um painel interativo utilizando a tecnologia Microsoft Power BI.

4.1. Princípios e Fontes

A elaboração do painel segue princípios de transparência, padronização visual, legibilidade, interatividade e orientação à tomada de decisão. A principal fonte de dados é o banco relacional do Observatório, abastecido pelo fluxo de coleta e tratamento da Produção Científica. O recorte de parceiros internacionais é aplicado sobre os registros já curados, evitando a criação de uma base paralela e garantindo consistência entre os indicadores de produção científica institucional.

4.2. Modelagem e Transformação

A modelagem relaciona os registros de publicações às informações de autores, unidades da Fiocruz, instituições parceiras, países, tipologias documentais, veículos de publicação, anos e áreas do conhecimento. Os identificadores persistentes gerados durante o tratamento permitem manter as relações entre as entidades e reduzir duplicidades. As transformações de maior complexidade são executadas antes da visualização, por meio dos scripts de tratamento e da estrutura do banco de dados; no Power BI, são organizadas as relações necessárias ao modelo analítico, os filtros do indicador e os elementos de apresentação.

4.3. Identidade Visual e Funcionalidade

A identidade visual do painel é alinhada às diretrizes institucionais da Fiocruz e do Observatório. O produto permite explorar as publicações com parceiros internacionais por meio de filtros e interações, possibilitando análises por período, unidade institucional, parceiro, país, tipo de produção, veículo e demais variáveis disponíveis no conjunto tratado. Sua disponibilização busca subsidiar análises estratégicas, ampliar a transparência e apoiar o acompanhamento da cooperação científica internacional da Fiocruz.

5. Análise Específica: Identificação de Parceiros Internacionais

A característica específica deste indicador é a seleção de publicações que apresentam colaboração internacional. Essa seleção depende da identificação e da harmonização das afiliações declaradas pelos autores, pois uma mesma instituição pode aparecer com abreviações, traduções, erros de grafia ou diferentes formas de escrita.

Inicialmente, as afiliações reconhecidas como pertencentes à Fiocruz são associadas às respectivas unidades institucionais. As demais afiliações são tratadas como potenciais instituições parceiras e passam por processos de padronização com dicionários, correspondência exata e aproximada, lógica fuzzy, distância de Levenshtein e curadoria manual ou semiautomatizada. Quando possível, os registros são enriquecidos com informações do ROR e outras fontes de referência, que auxiliam na identificação institucional e geográfica.

Depois da harmonização, as publicações são relacionadas às instituições e aos países identificados. O painel mantém apenas os registros que atendem ao critério de possuir ao menos um parceiro internacional, preservando simultaneamente o vínculo com a Fiocruz, a autoria, o ano, a tipologia documental e as demais informações analíticas. Dessa forma, o indicador representa um recorte do universo de produção científica institucional, e não uma coleta bibliográfica independente.

6. Considerações Técnicas

Esta seção detalha os principais desafios técnicos e os cuidados adotados para garantir a robustez, a rastreabilidade e a atualização do processo.

- **Heterogeneidade das fontes:** os registros são recebidos em diferentes formatos e estruturas. Scripts específicos convertem os dados para JSON, com harmonização de codificação UTF-8, correção de caracteres e remoção de elementos de formatação incompatíveis.
- **Variações de afiliação:** nomes institucionais podem conter abreviações, traduções e erros de digitação. A mitigação envolve dicionários construídos e revisados pela equipe, comparações exatas e aproximadas e curadoria com apoio do software VantagePoint.
- **Qualidade e completude:** identificadores como DOI, ISSN, ORCID e ROR nem sempre estão disponíveis. Por isso, o processo combina diferentes critérios de validação e fontes de enriquecimento, mantendo a rastreabilidade das bases de origem.
- **Controle de duplicidades:** a verificação do fiocruz_id, DOI, título, ano e veículo evita a multiplicação de uma mesma publicação quando ela é recuperada em mais de uma fonte.
- **Estratégia de atualização:** os dados armazenados passam por revisão manual semanal e recebem correções ou complementações de forma incremental. As melhorias incorporadas ao banco são refletidas no painel, cuja atualização também é programada com frequência semanal.
- **Padronização da saída:** os registros intermediários são mantidos em JSON e os dados consolidados são armazenados em PostgreSQL, favorecendo a integração com o Power BI e a manutenção de um fluxo reproduzível.

Referência metodológica complementar: [Metodologia do Painel de Indicadores de Produção Científica da Fiocruz](#). O documento apresenta o detalhamento integral da coleta, do tratamento, da harmonização, do armazenamento e da curadoria do conjunto de dados que dá origem a este indicador.