

Procedimento metodológico para coleta de dados e elaboração de painel de Projetos de Pesquisa Fiocruz na Plataforma Lattes

Data de Início do projeto: janeiro de 1962 a junho de 2026

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 Projetos de Pesquisa Fiocruz na Plataforma Lattes. 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 Projetos de Pesquisa Fiocruz na Plataforma Lattes no período de cobertura de dados informado.

2. Fontes de Dados

Esta seção descreve os sistemas e bases institucionais utilizados na identificação dos pesquisadores, na recuperação dos currículos e na delimitação dos vínculos com a Fiocruz.

2.1. Plataforma/Sistema

A principal fonte dos registros de projetos é a Plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Para identificar os currículos institucionais, foi utilizada a Base Lattes Fiocruz (BLF), sistema desenvolvido pela Coordenação-Geral de Gestão de Tecnologia de Informação (Cogetic), que importa e disponibiliza para consulta os currículos Lattes associados à instituição.

A relação de servidores e as informações sobre datas de entrada, saída e localização institucional foram fornecidas pela Coordenação-Geral de Gestão de Pessoas (Cogepe). Esses dados permitiram compatibilizar os períodos dos projetos com os períodos de vínculo institucional.

2.2. Conteúdo Agregado

A Plataforma Lattes agrega registros declarados pelos próprios pesquisadores. Para este indicador, foram recuperados todos os tipos de projetos existentes nos currículos: projetos de pesquisa, ensino, desenvolvimento, extensão e outros.

Entre os dados estruturados estão o ID Lattes do currículo, nome completo do autor principal, período de execução, título, natureza, situação, quantitativos de formação de alunos, descrição do projeto e informação sobre a existência de financiamento.

2.3. Acesso

Os IDs Lattes foram identificados por meio da Base Lattes Fiocruz. De posse desses identificadores, os currículos correspondentes foram coletados individualmente no formato XML. A extração concentrou-se nas informações sobre projetos cadastrados pelos servidores ativos e inativos presentes na relação institucional utilizada na coleta.

3. Procedimento Metodológico para Coleta e Tratamento

O processo foi estruturado para retirar a centralidade dos registros das pessoas e organizar a base em torno dos projetos desenvolvidos por servidores da Fiocruz. A Figura 1 apresenta o fluxo geral adotado, desde a identificação dos currículos até a consolidação dos projetos.

Identificação de IDs Lattes e Coleta de Currículos

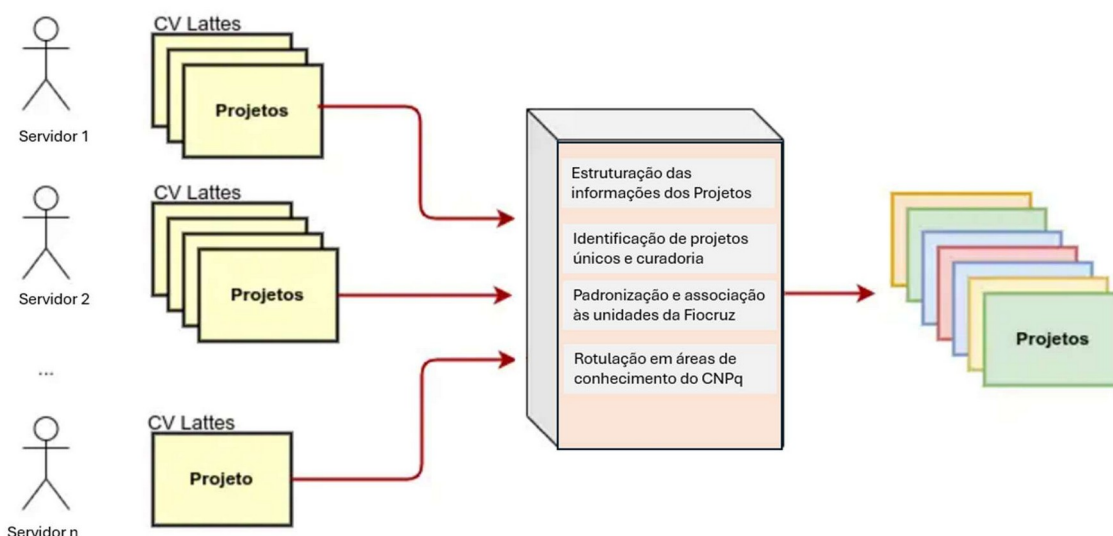


Figura 1. Esquema macro da coleta de dados a partir dos currículos Lattes.

3.1. Coleta dos Dados

3.1.1. Identificação dos IDs Lattes e coleta dos currículos

A partir da relação de servidores fornecida pela Cogepe, foram identificados na Base Lattes Fiocruz os IDs necessários à extração. Os currículos correspondentes foram obtidos em formato XML para cada servidor ativo e inativo incluído na relação.

Foram utilizadas as datas de entrada e eventual saída da instituição para compatibilizar a participação nos projetos com os períodos de vínculo institucional. Na contabilização, foram mantidos os períodos de sobreposição entre a execução do projeto e o vínculo efetivo do servidor com a Fiocruz.

3.1.2. Estruturação das informações dos projetos

Os registros extraídos dos currículos foram reorganizados em uma estrutura tabular no formato CSV, contendo os seguintes campos:

- ID Lattes do currículo;
- nome completo do autor principal do projeto;
- período de execução, com ano de início e ano de fim;

- título, natureza e situação do projeto;
- quantitativos de formação de alunos de graduação, especialização, mestrado e doutorado;
- descrição completa do projeto, quando disponível;
- informação sobre a existência ou não de financiamento.

3.2. Tratamento dos Dados

Como a Plataforma Lattes é de autopreenchimento, um mesmo projeto pode ser cadastrado por vários integrantes e apresentar diferenças de grafia, descrição, participantes ou datas. No currículo não há diferenciação padronizada entre coordenador, coordenador adjunto, aluno ou outra forma de participação. Por esse motivo, foi considerado integrante do projeto todo servidor que declarou participação no respectivo registro.

3.2.1. Filtragem Inicial e compatibilização do vínculo

A coleta incluiu currículos de servidores ativos e inativos. Para a contabilização dos projetos, as datas de execução foram comparadas às datas de entrada e saída dos servidores. Foram preservados os intervalos em que houve sobreposição entre o projeto e o vínculo institucional efetivo.

3.2.2. Identificação dos projetos únicos e deduplicação

Foi desenvolvido um script para identificar projetos cadastrados por mais de um integrante. Nos casos de duplicidade, foram mantidos como referência os registros com maior completude de descrições, títulos e listas de participantes.

Como as informações são digitadas livremente pelos pesquisadores, o documento de origem reconhece que não é possível assegurar a remoção integral de todas as repetições. Para ampliar a curadoria, foi realizada uma segunda rodada de limpeza no software VantagePoint®, com aplicação de algoritmos de lógica fuzzy para reunir títulos semelhantes.

3.2.3. Curadoria do período de execução

Quando projetos semelhantes ou idênticos apresentaram datas diferentes, foi adotado o período mais abrangente. Para o início, manteve-se o ano mais antigo; para o fim, o ano mais recente. Projetos com duração superior a 15 anos e sem ano de encerramento foram classificados com duração incompatível, considerando a possibilidade de desatualização do currículo.

3.2.4. Padronização e associação às unidades da Fiocruz

Após a deduplicação e a curadoria dos períodos, os pesquisadores foram associados às respectivas unidades da Fiocruz com auxílio da planilha institucional fornecida pela Cogepe. A unidade de localização atual do servidor foi adotada como referência.

3.2.5. Atribuição do rótulo de financiamento

Foi criado um indicador quantitativo para distinguir projetos financiados e não financiados, considerando fontes internas ou externas. A Plataforma Lattes não fornece, nesse campo, informações qualitativas sobre a fonte nem os valores financiados.

3.2.6. Rotulação em áreas do conhecimento

Os projetos foram classificados automaticamente em Grande Área, Área e Subárea do conhecimento, correspondentes aos três primeiros níveis da taxonomia do CNPq. A rotulação foi realizada por um modelo de Inteligência Artificial desenvolvido pelo Observatório CT&I em Saúde.

O modelo combina as técnicas TFIDF-LCPN-SVM e foi elaborado originalmente para a classificação de publicações científicas. Segundo a metodologia de origem, sua execução não exige hardware especializado, como unidades de processamento gráfico.

3.2.7. Armazenamento

Os registros estruturados foram armazenados em formato CSV, armazenados no SharePoint e consolidados em uma view utilizada na construção do painel.

4. Elaboração do Painel/Relatório

A partir da view consolidada, foi desenvolvido um painel interativo no Microsoft Power BI.

4.1. Princípios e Fontes

O painel utiliza os dados extraídos dos currículos Lattes, tratados e organizados com centralidade nos projetos. A visualização foi concebida para apresentar indicadores básicos sobre os diferentes tipos de projeto e sua evolução ao longo do tempo.

4.2. Modelagem e Transformação

A metodologia informa a construção de uma view consolidada para alimentar o Power BI. Entretanto, não detalha a quantidade de tabelas do modelo, as chaves de relacionamento, a cardinalidade, as etapas realizadas no Power Query ou as medidas DAX utilizadas.

4.3. Identidade Visual e Funcionalidade

O painel permite explorar a evolução dos projetos, analisar redes de colaboração entre pesquisadores, identificar as unidades mais atuantes e reconhecer pesquisadores centrais em termos de participação em projetos institucionais.

5. Análises Avançadas — Redes de Colaboração

A análise de redes foi utilizada para representar a coparticipação das unidades da Fiocruz em projetos. Cada vértice corresponde a uma unidade institucional e cada aresta conecta duas unidades que participaram de projetos em conjunto.

A espessura das arestas é proporcional ao número de projetos desenvolvidos em parceria: quanto mais espessa a ligação, maior o volume de colaboração entre as unidades. Essa abordagem permite observar a articulação entre diferentes grupos e identificar padrões de cooperação institucional.

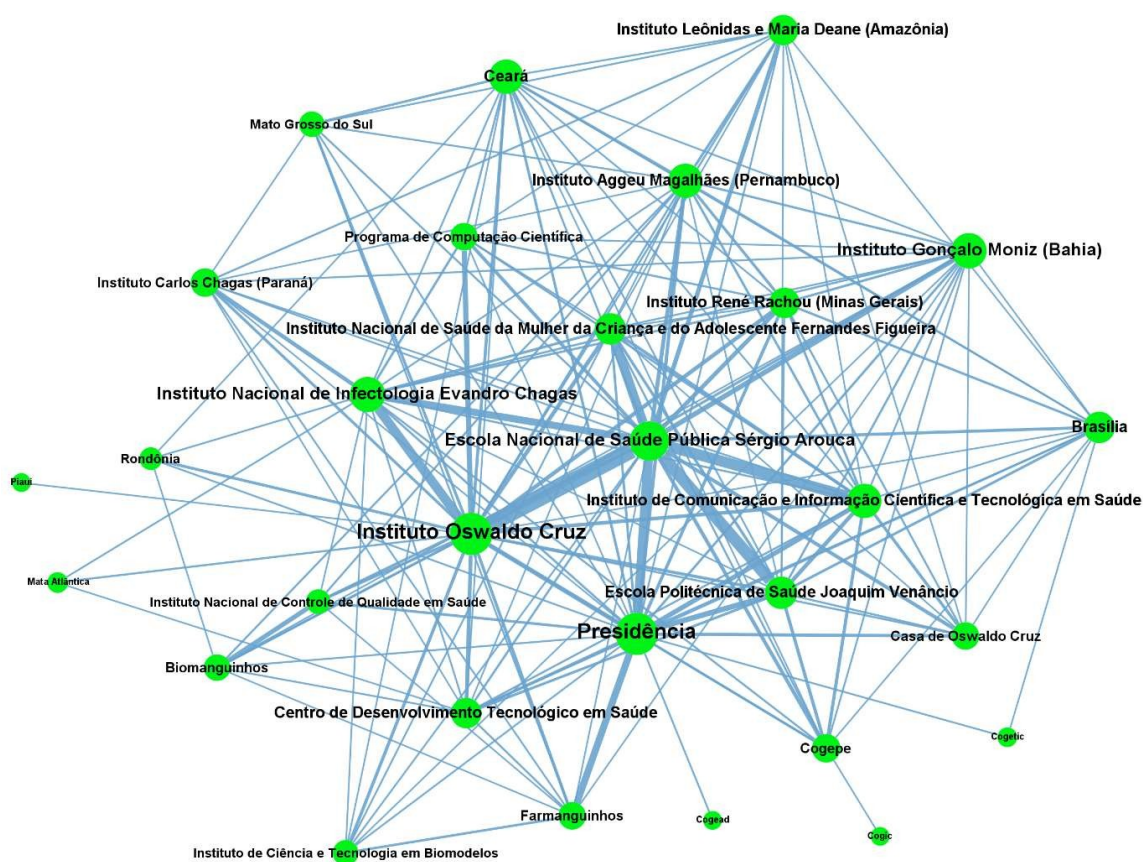


Figura 3. Rede de coparticipação em projetos. Cada vértice representa uma unidade.

6. Considerações Técnicas

A estratégia adotada oferece como principal vantagem a organização estruturada das informações registradas na Plataforma Lattes. Como os currículos são preenchidos pelos próprios pesquisadores, a atualização ocorre de forma descentralizada e a coleta pode ser automatizada em escala institucional.