

Procedimento metodológico para coleta de dados e elaboração de painel de Produção Científica da Fiocruz

Período de análise: janeiro de 2008 a julho de 2026

Fundação Oswaldo Cruz — Vice-Presidência de Educação, Informação e Comunicação

1. Tema e Contexto

1.1. Tema

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), cuja proposta é transformar dados provenientes de diversas fontes em informações com alto valor agregado para a Instituição e para a sociedade, por meio da construção de indicadores que favoreçam a tomada de decisão e proporcionem maior transparência à sociedade, fortalecendo direta e indiretamente as ações que compõem o Sistema Único de Saúde. Este documento descreve os procedimentos adotados na coleta, processamento, curadoria e disponibilização do indicador de produção científica da Fiocruz em um dashboard, formato de visualização que permite explorar os dados e analisar suas relações a partir de cliques e filtros. 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 a produção científica institucional no período de janeiro de 2008 a julho de 2026.

1.2. Contexto

Esta metodologia apresenta o processo completo em uma arquitetura em dois estágios com responsabilidades separadas e não sobrepostas: um pipeline de inserção que valida, converte, harmoniza e insere os dados das fontes no banco institucional ;e um pipeline de curadoria que audita e resolve anomalias e duplicidades acumuladas no banco com base em critérios pré-definidos,. A inserção é conservadora por princípio (não sobrescreve campo preenchido e não funde registros por semelhança de nome entre publicações distintas); a curadoria é a etapa que resolve essas pendências.

1.3. Objetivo do Trabalho e Escopo

Os dados de publicações científicas foram coletados, tratados, harmonizados, enriquecidos e inseridos em um banco de dados, a partir do qual foi construído um dashboard com indicadores de produção científica. O quadro a seguir resume o escopo.

ITEM	DESCRIÇÃO
Objeto	Produção científica de pesquisadores e servidores da Fiocruz
Fontes	7 bases bibliográficas + cadastro funcional (COGEPE) + Diretório de Grupos de Pesquisa do CNPq
Produto	Base relacional (SQL Server) com produções, pessoas, afiliações, veículos, palavras-chave e classificação em áreas do conhecimento do CNPq; dashboard interativo em Power BI
Unidade de análise	A produção científica (artigo, capítulo, tese etc.), identificada por DOI ou por hash de título + ano
Período	Janeiro de 2008 a julho de 2026
Fora do escopo	Cálculo de indicadores bibliométricos derivados (contagens, índices) — realizado na camada de BI, sobre o banco produzido por este processo

Princípio metodológico central da inserção: não descartar dado por falta de metadado e não sobrescrever informação já registrada (regra fill-only). Erros por inclusão são reversíveis; perda de dado e fusão indevida de identidades, não. O fechamento das duplicidades remanescentes é responsabilidade da etapa de curadoria (Seção 3.6), executada sob plano revisável e backup integral.

2. Fontes de Dados

2.1. Plataformas e Sistemas

A produção científica é recuperada em sete bases de dados: Web of Science (WoS), Scopus, PubMed, LILACS, SciELO, ARCA e Currículo Lattes. O escopo do trabalho também inclui o cadastro funcional da COGEPE e o Diretório de Grupos de Pesquisa do CNPq, conforme definido no escopo da metodologia.

ORDEM	BASE	FORMATO DE ENTRADA	NATUREZA DA COLETA
1ª	Web of Science (WoS)	XLSX	Exportação da busca institucional por afiliação
2ª	Scopus	CSV	Exportação da busca institucional por afiliação
3ª	PubMed	TXT (MEDLINE)	Exportação MEDLINE, múltiplos arquivos
4ª	LILACS	RIS	Exportação RIS da busca por afiliação

ORDEM	BASE	FORMATO DE ENTRADA	NATUREZA DA COLETA
5ª	SciELO	XML (JATS)	Coleção completa da rede, com pré-filtro por afiliação
6ª	ARCA	CSV / XLSX	Repositório institucional Fiocruz (parceria ICICT)
7ª	Lattes	XML por currículo	Currículos dos servidores (ativos e inativos)

2.2. Conteúdo Agregado e Fontes Complementares

FONTE	CONTEÚDO	USO NO PROCESSO
Cadastro COGEPE	Servidores Fiocruz: CPF, nome, unidade, vínculos	Principal identidade de autores e estabelecimento de vínculo do Lattes
OpenAlex [3]	Metadados por DOI: autores, ORCID, veículo, ISSN	Enriquecimento pós-parse e curadoria de DOIs/autores
Crossref [5]	Metadados de registro de DOI	Descoberta e validação de DOI; título canônico
ORCID [2]	Identificador persistente de autores e nomes públicos	Vinculação e verificação de identidade de autores
ROR	Identificador persistente de instituições	Enriquecimento institucional e guarda de deduplicação de afiliações
DOAJ	Metadados de periódicos de acesso aberto	Enriquecimento de veículos com ISSN cadastrado

2.3. Acesso e Forma de Obtenção dos Dados

A coleta ocorre em sete bases de dados: Web of Science (WoS), Scopus, PubMed, LILACS, SciELO, ARCA e Currículo Lattes. Nas bases WoS, Scopus, PubMed e LILACS, a coleta é realizada diretamente nos respectivos sites, utilizando uma string de busca por afiliação comum a todas as bases, com adaptação apenas de formato (Apêndice I). A string contempla de forma abrangente a heterogeneidade de formas de escrita da Fiocruz, suas unidades e escritórios regionais, incluindo os erros de grafia mais comuns.

A base SciELO é baixada integralmente em arquivos XML (coleção completa da rede), aplicando-se a cada arquivo um pré-filtro obrigatório com a mesma string de afiliação, o que reduz os cerca de 1 milhão de XMLs ao subconjunto pertinente. Os dados do ARCA, repositório institucional da Fiocruz, são disponibilizados pela equipe do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), sem necessidade de consulta por afiliação. Na Plataforma Lattes, são extraídos os currículos XML dos servidores ativos e inativos da Fiocruz, identificados a partir do cadastro funcional fornecido pela Coordenação-Geral de Gestão de Pessoas (COGEPE).

A ordem de inserção é canônica e não se inverte: WoS → Scopus → PubMed → LILACS → SciELO → ARCA → Lattes. Como a escrita é fill-only, a primeira fonte que grava um campo o congela; por isso as bases com metadado editorialmente controlado vêm primeiro e o Lattes, que é autodeclarado pelo titular, vem por último.

3. Procedimento Metodológico para Coleta e Tratamento

3.1. Coleta dos Dados

3.1.1. Visão Geral do Processo

O processo se organiza em cinco macroetapas encadeadas. Cada uma é descrita em detalhe nas seções seguintes.

#	MACROETAPA	DESCRIÇÃO RESUMIDA
1	Coleta	Busca por afiliação nas 7 bases bibliográficas, com string comum adaptada a cada interface (String de busca anexa); uso de dados de parceiros internos (COGEPE e ICICT)
2	Inserção	Ciclo validate → parse → enriquecimento → insert → update por fonte, com conversão das coletas em formato único JSON v2, harmonização por dicionários e algoritmos, e resolução de identidade de autores
3	Armazenamento	Persistência em banco relacional SQL Server via mapeamento objeto-relacional
4	Classificação e enriquecimento	Vinculação de áreas do conhecimento do CNPq (manual e por modelo de IA) e enriquecimento com ORCID, OpenAlex, DOAJ, ROR e Crossref
5	Curadoria da base	Auditoria e deduplicação global do acervo já carregado (nomes, DOIs, títulos, autores, afiliações, vínculos), com backup, plano prévio e rastreamento de critério

A execução da inserção é orquestrada por um runner próprio que executa os módulos em ordem canônica, registra o progresso e mantém log persistente com contexto por módulo e por registro. A curadoria é orquestrada separadamente, com modos de planejamento (dry-run) e aplicação (exec), backup automático e artefatos de revisão por execução.

3.1.2. Ciclo de Processamento por Fonte

Cada base está isolada em um módulo e todo módulo executa o mesmo ciclo, sem exceções:

#	FASE	O QUE FAZ	ACESSA BANCO?
1	validate	Confere existência e tamanho mínimo do arquivo-fonte	Não
2	parse	Lê a fonte e produz registros no formato JSON v2	Não
3	enriquecimento	Consulta o OpenAlex por DOI e complementa o registro	Somente rede
4	filtro de ano	Aplica o recorte temporal sobre o ano da produção	Não
5	insert	Persiste registros novos (idempotente)	Sim
6	update	Preenche campos vazios de registros existentes com dados complementares entre as bases	Sim

3.1.3. Contrato de Intercâmbio (JSON v2)

Cada registro produzido pelo parse é um documento JSON com blocos nomeados, validado por schema antes de qualquer escrita. O desenho separa as entidades dos relacionamentos entre elas: blocos de produção, fontes, veículos, palavras-chave, áreas, pessoas e afiliações alimentam as tabelas correspondentes, enquanto o bloco de autorias registra a ordem de autoria e as referências a pessoa e afiliação. Todo registro carrega um identificador persistente (hash SHA-1 de título + ano, o fiocruz_id), usado como chave de deduplicação quando não há DOI.

Especificamente para a Plataforma Lattes, que é orientada a autor, e não à publicação, um passo extra combina uma mesma publicação declarada em diferentes currículos antes da conversão, e considera as datas de admissão e eventual saída dos servidores para compatibilizar a produção com o período de vínculo institucional. Vínculo inativo sem data de desligamento mantém a janela aberta (nunca se descarta produção por falta de metadado) e é contabilizado e reportado para conserto na origem.

3.2. Tratamento dos Dados

Etapa crítica do processo, na qual são realizadas a limpeza, a homogeneização e a harmonização dos dados. Cada base fornece seus dados em estrutura e formato próprios, com graus distintos de completude, precisão e padronização. Após a coleta, os registros são convertidos para um formato único e submetidos a dicionários de referência e algoritmos de comparação exata, aproximada e apoiada por Inteligência Artificial.

3.2.1. Filtragem Inicial

Ao converter os registros, aplicam-se critérios de exclusão: registros sem título ou com título curto demais; sem ano de publicação ou com ano fora do intervalo janeiro de 2008 até julho de 2026; e sem autor. Também são descartadas as tipologias documentais que não configuram publicação inédita (errata, carta ao autor, comentário, correção, corrigendum, retratação, preprint, republicação), bem como publicações sem nenhum autor com afiliação Fiocruz identificada, critério este, que corrige falsos positivos clássicos da coleta, como a Escola Nacional de Saúde Pública de Lisboa, o Instituto Evandro Chagas do Pará e o Hospital Universitário Oswaldo Cruz de Recife.

3.2.2. Deduplicação

A deduplicação na inserção utiliza chaves específicas para cada entidade, priorizando identificadores persistentes e evitando fusões indevidas por mera semelhança de nome entre publicações distintas.

ENTIDADE	CHAVE PRIMÁRIA	CHAVE SECUNDÁRIA
Produção científica	DOI	Hash SHA-1 (título + ano)
Veículo	ISSN	Título
Afiliação	Nome harmonizado	—
Pessoa	CPF, ORCID ou id Lattes	Ids de plataforma → coautoria na mesma produção
Palavra-chave	(forma normalizada, tipo)	—
Área CNPq	(nível 1, 2, 3, 4)	—
Grupo de pesquisa	Nome do grupo	—

Ao inserir um registro, o pipeline verifica a existência prévia da publicação (por DOI ou hash de título + ano); se já presente, apenas a informação da base de origem é acrescida. Todo lookup de deduplicação é ordenado por identificador (menor vence), de modo que execuções distintas convirjam para a mesma linha.

3.2.3. Armazenamento

Um banco de dados relacional foi especialmente desenhado para abrigar estes dados, tendo como sistema gerenciador o Microsoft SQL Server (modelo de dados v3.1, com DDL canônico e dicionário de dados versionados). O mapeamento dos dados para o banco é realizado por mapeamento objeto-relacional (SQLAlchemy), com acesso parametrizado via driver ODBC. O ambiente de execução é disponibilizado em Docker, garantindo reprodutibilidade da infraestrutura.

As primeiras informações inseridas são as pessoas oriundas do cadastro COGEPE, com suas unidades de afiliação, CPF, identificador Lattes e ORCID (Seção 3.4.1). Em seguida, são armazenados os dados das publicações, obedecendo à ordem de precedência por qualidade e completude dos campos: 1ª WoS; 2ª Scopus; 3ª PubMed; 4ª LILACS; 5ª SciELO; 6ª ARCA; 7ª Plataforma Lattes. A escrita fill-only garante que a informação da fonte de maior qualidade editorial prevaleça, e a idempotência garante que reexecutar a mesma fonte não duplique registros.

3.3. Harmonização dos Dados

3.3.1. Harmonização de Afiliações - Cascata de Seis Etapas

Para cada string de afiliação encontrada nos registros, aplica-se uma sequência que para na primeira etapa que resolver:

ETAPA	MÉTODO	RESULTADO
1	Expressão regular por unidade Fiocruz (inclui aliases geográficos, p.ex. "Fiocruz Minas" → Instituto René Rachou)	Unidade Fiocruz determinística
2	Predicado "é Fiocruz" + dicionário de unidades (comparação exata e aproximada), com IA generativa (Maritaca sabiá-4-thinking, com fallback GPT-4o-mini) de apoio	Unidade Fiocruz
3	Dicionário de unidades Fiocruz, comparação aproximada (casos residuais)	Unidade Fiocruz
4	Dicionário de instituições parceiras, comparação exata e aproximada	Instituição parceira
5	Identificação de instituições parceiras com IA generativa (Maritaca sabiá-4-thinking, com fallback GPT-4o-mini) para o nome institucional canônico	Instituição parceira

ETAPA	MÉTODO	RESULTADO
6	API ROR: nome em inglês, sigla e identificador ROR (etapa opcional, desligada por padrão)	Enriquecimento institucional da instituição parceira

Falha externa nunca vira dado: indisponibilidade de rede ou de API é distinguida de resposta negativa, e apenas respostas efetivas são registradas e cacheadas. Os dicionários de referência foram construídos ao longo dos últimos seis anos por especialistas do Observatório, com auxílio do software VantagePoint® e revisão manual, e permanecem como importante insumo para a identificação das unidades Fiocruz, parceiros, tipologia documental, veículos, ISSN e países.

3.3.2. Demais Harmonizações

OBJETO	PROCEDIMENTO	FALLBACK
Veículo de publicação	Comparação por ISSN no dicionário; na ausência, por título	Título original da fonte
ISSN	Normalização para o padrão NNNN-NNNC	Campo vazio + advertência
Tipo documental	Dicionário de tipos → identificador harmonizado	Tipo "não identificado"
Idioma	Normalização de rótulo de idioma	Campo vazio
Palavras-chave	Chave (forma normalizada, tipo); reaproveita variantes	Nova palavra-chave
Ano e tipo	Conversão na fronteira de escrita; faixa de ano válida	Campo vazio + advertência
Nome de pessoa	Normalização de acentos, partículas e ordem de citação	Nome original preservado

As comparações aproximadas utilizam lógica fuzzy e distância de Levenshtein [4]. No caso dos nomes de autores, um procedimento específico homogeneiza variações de citação, abreviações e inversões (p.ex. "Silva, BC", "B C Silva" e "Silva B. C." convergem para "B. C. Silva"), preservando sempre o nome original da fonte em coluna própria (person_original_name).

Eventos acadêmicos declarados como veículo — sobretudo na Plataforma Lattes — são harmonizados sob nome guarda-chuva, agrupando edições e variações de grafia (p.ex. as edições XXI, XXII e XXIII do Congresso da Sociedade Brasileira de Parasitologia sob "Congresso da Sociedade Brasileira de Parasitologia"), sem prejuízo de identificação da edição, exibida junto ao ano de publicação.

3.4. Identidade de Autores

3.4.1. Pré-passo de Identidade (Servidores)

Antes de qualquer fonte bibliográfica, um pré-passo insere/atualiza o cadastro de pessoas com identidade verificada, para que os autores encontrados nas bases sejam unificados com a pessoa já existente em vez de criar duplicatas. A etapa 1 combina o cadastro COGEPE, o mapa CPF → identificador Lattes e o ORCID extraído de cada currículo XML, gerando um arquivo com os dados pessoais cifrados para garantir conformidade às regras de tratamento de dados pessoais da Fiocruz; a etapa 2 lê esse arquivo e faz a carga das pessoas e de seus vínculos de afiliação. Nenhum parser bibliográfico emite CPF ou identificador Lattes (exceto o do próprio Lattes) e a ponte entre o cadastro funcional atualmente usada para os autores das bases é o ORCID, cuja cobertura é medida e registrada em toda execução.

3.4.2. Escada de Identidade de Pessoas

NÍVEL	EVIDÊNCIA	ESCOPO	EFEITO
Tier 1	CPF, ORCID ou identificador Lattes	Global	Reaproveita a pessoa existente
Tier 2	Identificadores de plataforma (WoS, Scopus, PubMed)	Global	Reaproveita a pessoa existente
Tier 2.5	Compatibilidade de nome entre coautores da mesma produção	Uma publicação	Reaproveita a pessoa existente
—	Nenhuma das anteriores	—	Cria uma nova pessoa

Regra de unificação por nome: semelhança de nome só resolve identidade dentro de uma mesma publicação, onde dois homônimos distintos são improváveis. Entre publicações distintas, apenas identificador persistente unifica registros de pessoas. Autor sem identificador persistente gera uma pessoa por publicação, com duplicação consolidável a posteriori pela curadoria. Isso evita a fusão indevida, que não tem rollback.

3.4.3. Curadoria de Identidade Opcional na Inserção

PROCEDIMENTO	O QUE FAZ	SALVAGUARDA
Canonicalização de ORCID	Reescreve grafias legadas para XXXX-XXXX-XXXX-XXXX	Só reescreve texto que já é puramente um ORCID
Verificação de ORCID	Compara o nome do banco com o registro público ORCID e zera identificadores sem correspondência	Zera só com evidência; falha de rede nunca zera; teto de 5% recusa zeramento em massa

PROCEDIMENTO	O QUE FAZ	SALVAGUARDA
Unificação por CPF	Funde linhas de pessoa duplicadas ancoradas em CPF	Destrutivo; nunca roda na execução padrão; auditoria completa em log
Relatório de vínculos	Lista vínculos inativos sem data de desligamento	CPF anonimizado

3.5. Isolamento de Falhas (Quarentena)

O método assume que fontes reais contêm registros anômalos. Toda falha é isolada por registro e nunca por arquivo ou por módulo. XML malformatado é quarentenado sem interromper os demais; erro de banco provoca rollback apenas do registro afetado; deadlocks são retentados com backoff antes da quarentena. Registros quarentenados podem ser reprocessados isoladamente após a correção da causa, sem reparsear a fonte inteira.

3.6. Curadoria da Base de Dados

A curadoria opera sobre o banco já carregado, em ciclo read → plan → review → apply: lê o estado atual, gera um plano de mudanças em artefatos de revisão (CSV/JSONL), submete o plano à revisão e só então aplica. Nenhum módulo escreve sem plano prévio, e toda aplicação é precedida de backup CSV completo de todas as tabelas, respeitado o ambiente em uso (DEV, TST, HOM, PRD), de modo que um backup de homologação nunca sirva de lastro para uma escrita em produção.

3.6.1. Fronteira entre Inserção e Curadoria

DIMENSÃO	INSERÇÃO (PIPELINE)	CURADORIA (AUDITORIA)
Entrada	Arquivos das fontes	O próprio banco já carregado
Ciclo	validate → parse → insert → update	read → plan → review → apply
Escopo de decisão	Dentro de uma fonte e de uma publicação	Global, em todo o acervo
Postura diante do dado	Não substitui campo preenchido	Substitui, funde e apaga — com plano revisado antes
Resolução de pessoa	Identificador persistente; nome só dentro da publicação	Identificador forte canônico + similaridade de nome sob portões explícitos
Reversibilidade	Reprocessamento da fonte	Backup CSV completo antes de qualquer escrita

Regra de fronteira: tudo que exige comparar registros de publicações diferentes, apagar linha ou reescrever campo já gravado é curadoria. Tudo que depende do arquivo-fonte original é inserção.

3.6.2. Procedimentos de Curadoria

ETAPA	OBJETIVO	CHAVE DE DECISÃO
Normalização de nomes	Padronizar nomes de pessoa (pontuação, ordem de citação, abreviações, campo vazio)	Regras determinísticas de texto, sem consulta externa
Enriquecimento de DOI	Gravar DOI ausente, validar DOI existente e alinhar títulos	Similaridade de título $\geq 0,99$ com confirmação reversa em Crossref [5] e OpenAlex [3]; conflito entre fontes não grava
Reconciliação de autores	Reconciliar autores de publicações com DOI contra o OpenAlex	ORCID exato ou similaridade de nome $\geq 0,85$ com sobrenome compatível; afiliação Fiocruz curada nunca é sobrescrita
Dedup intra-publicação	Fundir autores duplicados dentro da mesma publicação	Score composto de nome e afiliação $\geq 0,85$, ou CPF/ORCID canônicos iguais; identificadores divergentes bloqueiam
Consolidação de keywords	Unificar variantes de descritores	Comparação exata normalizada e catálogos curados (países, saúde); tipos distintos nunca se fundem
Catálogo de áreas CNPq	Alinhar o catálogo de áreas à tabela oficial vigente	DE/PARA por código e nome
Unificação global de pessoas	Fundir duplicatas de pessoa em todo o banco	Identificador forte exato ancorado em CPF; nunca por similaridade de nome
Harmonização de títulos	Recuperar títulos corrompidos por codificação de caracteres	Substituição do título inteiro, com âncora de identidade (DOI ou vencedor único) e preservação do original
Dedup de autorias	Remover vínculos autor-publicação duplicados	Tupla idêntica (produção, pessoa, afiliação, ordem); vencedor é a linha mais completa
Dedup de afiliações	Unificar afiliações de nome idêntico	Nome exato, com bloqueio quando identificadores ROR divergem; Fiocruz nunca perde a eleição

3.6.3. Regras Transversais de Decisão

- Identificador canônico: toda decisão (agrupar, eleger, herdar, bloquear, forçar fusão) usa a forma canônica dos identificadores (CPF, ORCID, Lattes), que rejeita valores vazios, malformados e placeholders. Placeholder não é identificador: não é herdado, não gera conflito e não força fusão.
- Precedência: CPF > Lattes > ORCID. CPF e ORCID coincidentes forçam fusão; Lattes bloqueia divergência, mas não força fusão — decisão registrada para não criar fusões novas sem âncora forte.

- Assimetria entre bloquear e fundir: bloquear uma fusão aceita evidência frouxa (falso positivo apenas impede uma fusão); fundir exige forma canônica estrita (falso positivo causa uma fusão indevida). A direção da falha determina o rigor.
- Curadoria manual soberana: afiliações Fiocruz curadas manualmente nunca são sobrescritas e nunca perdem eleições de deduplicação.
- Rastro de critério: todo valor curado carrega, em coluna própria, o critério que o produziu; nomes, títulos e afiliações originais são preservados em colunas dedicadas — nada se perde.

3.6.4. Governança da Execução

MECANISMO	DESCRIÇÃO
Planejar antes de aplicar	Modo dry-run gera artefatos de revisão sem escrever no banco; modo exec aplica com backup prévio
Promoção entre ambientes	Teste → homologação → produção, revisando os artefatos de dry-run entre as etapas; escrita em produção exige confirmação explícita
Rastreabilidade	Identificador de execução carimbado em logs, sumários e resultados; todo sumário grava o bloco de conexão (ambiente, host, base) como procedência
Proteção de dados pessoais	CPF e e-mail mascarados nos artefatos de revisão; credenciais somente em arquivo de ambiente; artefatos fora do versionamento
Verificação	Suíte de testes por módulo, verificação offline do catálogo de servidores (nenhum nome resolve para o CPF de outra pessoa), execução-amostra e comparação read-only entre ambientes

4. Elaboração do Painel/Relatório

4.1. Princípios e Fontes

O painel é construído a partir dos dados armazenados no banco institucional produzido pelo processo metodológico. O cálculo de indicadores bibliométricos derivados, como contagens e índices, é realizado na camada de BI sobre o banco produzido por este processo.

4.2. Funcionalidade e Atualização

O dashboard permite explorar os dados e analisar suas relações a partir de cliques e filtros.

A partir dos dados armazenados no banco institucional, o painel interativo que esta metodologia acompanha é construído com a tecnologia Microsoft Power BI. Os dados do banco passam por revisão contínua, de forma que melhorias advindas de correções, complementação de dados e novos ciclos de curadoria são incrementalmente incorporadas. O dashboard está programado para atualizar-se a partir do banco de dados com frequência semanal.

5. Análises Avançadas - Classificação e Enriquecimento

5.1. Classificação em Áreas do Conhecimento do CNPq

A categorização de cada publicação por áreas do conhecimento do CNPq consiste em identificar e vincular à produção científica da Fiocruz às áreas do CNPq como parâmetro para análises por parte da Fiocruz sobre os temas em que mais tem trabalhado. Adota-se a classificação hierárquica do CNPq em quatro níveis: Grande Área, Área, Subárea e Especialidade, a qual é amplamente conhecida e utilizada pelos pesquisadores do país. A informação é obtida de duas formas:

- Diretamente das unidades ou da Plataforma Lattes: representantes das unidades da Fiocruz categorizaram amostras de suas publicações, e classificações registradas na própria Plataforma Lattes também foram consideradas; parceria com a Vice-Presidência de Pesquisa e Coleções Biológicas (VPPCB). Essa via responde por cerca de 55% do total das produções classificadas.
- Modelo de Inteligência Artificial: O Observatório desenvolveu um modelo próprio de Inteligência Artificial capaz de rotular as produções automaticamente. Os detalhes do modelo estão publicados em artigo científico [1].

5.1.1. Modelo de Classificação

ITEM	ESPECIFICAÇÃO
Arquitetura	Classificação hierárquica Local Classifier per Node (LCPN) sobre TF-IDF + SVM
Saída	4 rótulos por produção: Grande Área → Área → Subárea → Especialidade
Idioma de treino	Inglês
Pré-processamento	Remoção de dígitos/URLs/tags, caixa baixa, tokenização e stemming (reproduzidos exatamente do treino)

ITEM	ESPECIFICAÇÃO
Reprodutibilidade	Versões das bibliotecas de aprendizado fixadas — desserializar o modelo em outra versão pode divergir em silêncio

5.1.2. Etapas do Procedimento de Classificação

#	ETAPA	DESCRIÇÃO E SALVAGUARDA
1	Fila	Seleciona produções sem classificação vigente deste classificador; classificação em área aposentada não conta como classificada — a fila se autocura
2	Montagem do texto	Concatena título, resumo, palavras-chave, áreas do veículo e nome do veículo — a mesma receita usada no treino, versionada
3	Tradução	Detecta o idioma e verte o texto para inglês; falha de tradução deixa a produção sem classificar (volta na execução seguinte) — nunca se classifica o texto original
4	Predição	O modelo devolve os 4 rótulos, com correspondência estrita entre textos e produções
5	DE/PARA de áreas aposentadas	Cascata: caminho vigente → remapeamento curado → poda do nível mais profundo → não classifica; cobre 100% dos caminhos que o modelo pode emitir
6	Gravação	Resolve o identificador de área contra o banco de destino, pelos rótulos dos 4 níveis; recusa gravar em área aposentada
7	Checkpoint	Registra a decisão em cache de disco, permitindo interromper e continuar

A tradução é pré-requisito, não acessório: medição direta com o mesmo conteúdo mostrou que a resposta do modelo muda com o idioma do texto submetido. Por isso a tradução tem cache próprio e sua falha interrompe a classificação daquela produção.

5.1.3. Tratamento da Árvore CNPq e Revisão de Hierarquia

Para ambas as vias de rotulação (manual ou automatizada), realiza-se revisão de adequação das áreas dentro da hierarquia definida pelo CNPq, corrigindo erros de indentação de níveis (p.ex. Antropologia registrada como Grande Área, quando é Área dentro de Ciências Humanas). Áreas "novas" sugeridas pelas unidades (Bioinformática,

Tecnologias Assistivas, Direito Sanitário, entre outras) não são consideradas a priori no dashboard, e permanecem registradas como sugestão de ampliação da tabela do CNPq.

5.2. Enriquecimento dos Dados

Para cada produção cujo veículo de publicação tem ISSN cadastrado no DOAJ, os metadados correspondentes são importados para o banco de dados. Os identificadores ORCID [2], para autores, e ROR, para instituições, são vinculados a cada registro sempre que o vínculo pode ser estabelecido. A base OpenAlex [3] é utilizada para harmonizar nomes de veículos de publicação e, ao final do parse de cada fonte, para complementar os registros por DOI (ORCID de autores, ISSN e título do veículo, identificador OpenAlex), com critério conservador de casamento de autor: ORCID coincidente, ou similaridade de nome acima de limiar com sobrenome idêntico — empate não resolve. Todas as consultas externas são guardadas em disco, persistentes entre execuções, com escrita atômica e registro de respostas negativas (não repetir consulta sem resultado).

6. Considerações Técnicas

6.1. Qualidade e Reprodutibilidade

MECANISMO	DESCRIÇÃO
Idempotência	Reexecutar a mesma fonte não duplica registros; a segunda execução reporta zero inserções
Regra fill-only	Campo preenchido nunca é sobrescrito na inserção; exceções são explícitas e escopadas
Validação prévia	Arquivos auxiliares (dicionários) conferidos antes de qualquer módulo rodar, com falha imediata
Execução única	Trava de arquivo garante um único processo escrevendo no banco por vez
Monitoramento	Heartbeat de progresso por fase e por registro, com detecção de travamento; log persistente e rotativo
Quarentena	Falha isolada por registro em arquivo próprio, reprocessável
Tetos que se anunciam	Todo limite que pode descartar dado emite advertência ao ser atingido
Testes automatizados	Suítes executáveis sem banco, com casos de regressão para cada armadilha já observada
Modo de simulação	Dry-run executa validação, parse e análise sem escrever no banco

6.2. Governança e Proteção de Dados Pessoais

O tratamento de dados pessoais observa a Lei Geral de Proteção de Dados (LGPD). O cadastro funcional circula exclusivamente com CPF, data de nascimento e demografia

cifrados (criptografia simétrica com chave dedicada); o arquivo com CPF em claro é insumo exclusivo da etapa de geração e está excluído do ambiente de execução. CPF é sempre mascarado em logs e artefatos de revisão; nome civil e dados demográficos não entram em log. Identificadores públicos e autodeclarados (ORCID, id Lattes) podem constar em claro. Credenciais de acesso residem apenas em arquivo de configuração de ambiente — nunca em código, linha de comando ou artefatos.

6.3. Limitações Conhecidas e Decisões Assumidas

As limitações a seguir são registradas explicitamente porque não constituem defeitos: são fronteiras do método, com a decisão adotada e sua justificativa.

LIMITAÇÃO	DECISÃO ADOTADA	RAZÃO
Autor sem identificador persistente gera uma pessoa por publicação	Aceita; consolidação a posteriori pela curadoria	Fusão indevida não tem desfazer; duplicação é consolidável
Vínculos inativos sem data de saída mantêm a janela de coleta aberta	Aceita e reportada para conserto na origem	Descartar produção verdadeira é pior que incluir produção a mais
Iniciais comprimidas em nomes não são expandidas	Aceita	Distinguir duas iniciais de um sobrenome curto exigiria adivinhar
Modelo de áreas treinado na árvore antiga do CNPq	DE/PARA obrigatório por rótulo	Parte dos caminhos previstos só existe como área aposentada
Produções agrupadas sob o rótulo genérico "Fiocruz"	Mantidas até identificação da unidade	Padronização de nomes institucionais é problema sem solução totalmente eficaz; recomenda-se padronizar a declaração de afiliação
Cobertura de CPF limita a unificação global de pessoas por identificadores secundários	Âncora de CPF não relaxada; caminho de destrave é o preenchimento assistido pelo cadastro funcional	Fundir por ORCID sem âncora de CPF está fora de escopo por segurança
Enriquecimento institucional via ROR desligado por padrão	Opcional, por execução	Latência e quota de API externa

Referências

- [1] Hierarchical Article Classification: A Multi-Level Framework for Organizing Scholarly Literature. IEEE Access, 2025. DOI: 10.1109/ACCESS.2025.3579232.
- [2] Artigo sobre o identificador ORCID. Informação na Sociedade Contemporânea, v. 9, n. 2, 2023. DOI: 10.21680/2447-7842.2023v9n2ID33661.
- [3] PRIEM, J.; PIWOWAR, H.; ORR, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833, 2022.

[4] LEVENSHTAIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, v. 10, n. 8, p. 707–710, 1966.

[5] HENDRICKS, G. et al. Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies, v. 1, n. 1, p. 414–427, 2020. DOI: 10.1162/qss_a_00022.

Apêndice I - String de Busca nas Bases de Dados

String de busca por afiliação institucional, comum às bases WoS, Scopus, PubMed, LILACS e SciELO, com adaptação apenas de formato à interface de cada base.

("Biblioteca Manguinhos" OR "Bio Manguinhos" OR "Biomanguinho" OR "Biomanguinhos" OR "Inst Tecnol Immunobiol Biomanguinhos" OR "Inst Tecnol Immunobiol/Fiocruz" OR "Inst Tecnol Immunobiol-Fiocruz" OR "Bio Manguinhos" OR "Biomanguinhos" OR "Bio-Manguinhos" OR "Biomanguinlios" OR "Inst Tecnol Immunobiol" OR "Instituto de Tecnologia em Immunobiológicos" OR "Institute of Technology in Immunobiologicals" OR "Technology in Immunobiologicals Institute" OR "Ctr Desenvolvimento Tecnol Saude" OR "Ctr Technol Dev Health Fiocruz" OR "Ctr Technol Dev Hlth CdtS" OR "Ctr Tecnol Oswaldo Cruz" OR "CDTS/Fiocruz" OR "CDTSFiocruz" OR "Centro de Desenvolvimento Tecnológico em Saúde" OR "Center for Technological Development in Health" OR "Centro de Desenvolvimento Tecnológico Em Saúde" OR "Cent de Desenvolvimento Tecnológico em Saúde" OR "Centr de Desenvolvimento Tecnológico em Saúde" OR "Technological Development in Health Center" OR "Casa de Oswaldo Cruz" OR "Casa de Oswaldo Cruz" OR "Casa de Oswaldo Crus" OR "House of Oswaldo Cruz" OR "Oswaldo Cruz's House" OR "COC/Fiocruz" OR "COCFiocruz" OR "CRIS/Fiocruz" OR "CRISFiocruz" OR "Centro de Relações Internacionais em Saúde" OR "Center for International Health Relations" OR "Escola Nacional de Saúde Pública Sérgio Arouca" OR "Brazilian Natl Sch Publ Hlth" OR "Brazilian National School of Public Health" OR "Escola Nacl Saude Publ/Ensp" OR "Escola Nacl Saude Publ-Ensp" OR "Escola Nacl Saude Publ/Fiocruz" OR "Escola Nacl Saude PublFiocruz" OR "Escola Nacl Saude Publ Sergio Arouca" OR "Escola Natl Saude Publ/Fiocruz" OR "Escola Natl Saude Publ-Fiocruz" OR "Escuela Nacl Salud Publ/Fiocruz" OR "Escuela Nacl Salud PublFiocruz" OR "Escola Nacional de Saude Publica" OR "Escola Nacl Saude Publ" OR "Escola Nacl Saude Publica" OR "Escola Natl Saude Publ" OR "Natl Sch Publ Hlth/Fiocruz" OR "Natl Sch Publ Hlth-Fiocruz" OR "Natl Sch Publ Hlth Oswaldo Cruz Fdn" OR "Escola Politécnica de Saúde Joaquim Venâncio" OR "EPSJV" OR "Instituto de Tecnologia em Fármacos" OR "Pharmaceuticals Technology Institute/Fiocruz" OR "Pharmaceuticals Technology Institute-Fiocruz" OR "Far Manguinhos" OR "Farmanguinhos" OR "Far-manguinhos" OR "Inst Tecnol Farm-Fiocruz" OR "Inst Tecnol Farmacos/Fiocruz" OR "Inst Tecnol Farmacos-Fiocruz" OR "Inst Tecnol Farmacos Manguinhos" OR "Avenida Brasil 4365" OR "Avenida Brasil 4360" OR "Av Brasil 4365" OR "Av Brasil 4360" OR "Fdn Inst Oswaldo Cruz" OR "Fdn Oswald Cruz" OR "Fundacao Oswaldo Cruz" OR "Fundação Oswaldo Crus" OR "Fundação Oswaldo Cruz" OR "Fundação Osaldo Cruz" OR "Fundação Osawaldo Cruz" OR "Fundação Osqaldo Cruz" OR "Fundação Oswaldo Cruz" OR "Fundação Oswald Cruz" OR "Fundação Oswaldo Crus" OR "Fundação Oswaldo Cruzs" OR "Fundação Oswaldo Curz" OR "Fundação OSWALDO CURZ" OR "Fundação Oswalo Cruz" OR "Fundação Owaldo Cruz" OR "Fdn Oswaldo Crus" OR "Fdn Oswaldo Cruz" OR "Fdn Osaldo Cruz" OR "Fdn Osawaldo Cruz" OR "Fdn Osqaldo Cruz" OR "Fdn Oswaldo Cruz" OR "Fdn Oswald Cruz" OR "Fdn Oswaldo Crus" OR "Fdn Oswaldo Cruzs" OR "Fdn Oswaldo Curz" OR "Fdn OSWALDO CURZ" OR "Fdn Oswalo Cruz" OR "Fdn Owaldo Cruz" OR "FICORUZ" OR "FIO CRUZ" OR "FIOCRUZ" OR "FIOCUZ" OR "FIOCRUZ" OR "Fiocruz" OR "Oswaldo Crus Foundation" OR "Oswaldo Cruz Foundation" OR "Osaldo Cruz Foundation" OR "Oswaldo Cruz Foundation" OR "Osqaldo Cruz Foundation" OR "Oswaldo Cruz Foundation" OR "Oswald Cruz Foundation" OR "Oswaldo Crus Foundation" OR "Oswaldo Cruzs Foundation" OR "Oswaldo Curz Foundation" OR "OSWALDO CURZ Foundation" OR "Oswalo Cruz Foundation" OR "Owaldo Cruz Foundation" OR "Centro de Estudos Estratégicos/Fiocruz" OR "Centro de Estudos EstratégicosFiocruz" OR "CEE/Fiocruz" OR "CEE-Fiocruz" OR "VPAAPS/Fiocruz" OR "VPAAPSFiocruz" OR "VPEIC/Fiocruz" OR "VPPCB-Fiocruz" OR "VPPCB/Fiocruz" OR "VPPCBFiocruz" OR "VPPIS/Fiocruz" OR "VPPIS-Fiocruz" OR "Leonidas Maria Deane" OR "Leônidas e Maria Deane" OR "Leônidas & Maria Deane" OR "Leonidas Maria Deanne" OR "Leônidas e Maria Deanne" OR "Leônidas & Maria Deanne" OR "FiocruzAm" OR "Fiocruz/Am" OR "Fiocruz-Amasonia" OR "Fiocruz/Amasonia" OR "FiocruzAmazon" OR "Fiocruz/Amazon" OR "Fiocruz/Amazonas" OR "Fiocruz-Amazonas" OR

descritores da curadoria (Seção 3.6.2), com catálogos curados de países e de termos de saúde.

VARIANTES (KEY)	FORMA CANÔNICA (VALUE)
Adolescente / Adolescent	Adolescente
Amazônia / Amazon	Amazônia
Atenção Primária à Saúde / Atenção primária à saúde	Atenção primária à saúde
Brazil / Brasil	Brasil
COVID-19 / Covid-19	Covid-19
Children / Child	Criança
Diagnóstico / Diagnosis	Diagnóstico
Epidemiology / Epidemiologia	Epidemiologia
Inflammation / Inflamação	Inflamação
Malária / Malaria	Malária
Mortality / Mortalidade	Mortalidade
Prevalência / Prevalence	Prevalência
Primary Health Care / Primary health care	Primary health care
Saúde do trabalhador / Saúde do Trabalhador	Saúde do trabalhador
Saúde pública / Saúde Pública / Public health	Saúde pública
Unified Health System / SUS / Sistema Único de Saúde	Sistema Único de Saúde
Tuberculosis / Tuberculose	Tuberculose
Violência / Violence	Violência
Zika virus / Zika	Zika