Fundação Oswaldo Cruz - Fiocruz

Metodologia

Painel de indicadores de produção científica

Contextualização

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 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 institucional no período de janeiro de 2008 até junho de 2025.

Visão geral da metodologia

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. A **Figura *1*** apresenta uma visão geral das etapas do processo que é descrito em detalhes nesta metodologia.

Figura1: Etapas gerais do processo de elaboração do indicador da produção científica da Fiocruz.

A **coleta** ocorreu em sete bases de dados: Web of Science (WoS), Scopus, PubMed, Lilacs, SciELO, Arca e Currículo Lattes. Nas bases WoS, Scopus, PubMed e Lilacs, foi utilizada uma *string* de busca por afiliação, comum a todas as bases, apenas com adaptação no formato. A base SciELO foi completamente baixada em arquivos no formato XML, aplicando a cada arquivo a mesma *string* de busca de afiliação. Os dados do Arca foram disponibilizados pela equipe do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) da Fiocruz em formato Excel. Na Plataforma Lattes foram extraídas informações dos currículos dos servidores ativos e inativos da Fiocruz com base em seu respectivo CPF. Adicionalmente, foram coletados dados complementares de bases de

apoio, como ORCID, OpenAlex, DOAJ e CrossRef, para auxiliar no estabelecimento de vínculos entre os dados e melhorar a completude da informação coleta.

O tratamento e harmonização dos dados envolveu diversas etapas organizadas em um *pipeline*[1]. Cada coleta tem uma estrutura e um formato próprio, com maior ou menor grau de completude, precisão e padronização. Por isso, após a coleta, os dados foram tratados para garantir uma homogeneização, removendo publicações fora do intervalo pré-estabelecido, publicações do Lattes em intervalos não relacionadas com a permanência dos servidores na Fiocruz, publicações onde não há autores declaradamente vinculados à Fiocruz ou de tipos como errata, *re*spostas de autores, comentários e correções, entre outros, mantendo apenas publicações inéditas. Uma vez filtradas, as publicações foram convertidas para um formato único, o JSON (*JavaScript Object Notation*), considerando um mapeamento entre os campos do formato original da base e os campos definidos no arquivo JSON.

A partir do conjunto de todas as publicações de todas as sete bases convertidas em JSON, os dados passaram então por uma **harmonização**, feita a partir de dicionários e algoritmos. Estes são capazes de padronizar algumas informações, identificar irregularidades e corrigi-las ou sinalizá-las para correção manual ou com auxílio do *software* comercial Vantage Point®. A harmonização a partir de dicionários baseia-se em comparações exatas e aproximadas, que variam de 95 a 98% de *match* entre as *strings* com vistas a identificar dados iguais que foram escritos de forma distinta. Uma vez harmonizados, cada publicação é salva em uma segunda versão de arquivo JSON contendo os dados prontos para serem inseridos no banco de dados.

A inserção dos dados é feita a partir de um mapeamento objeto-relacional, onde o conteúdo do arquivo JSON harmonizado é mapeado para as tabelas do banco de dados do Observatório. Este banco de dados contém dados previamente inseridos com os identificadores únicos ORCID (para servidores da Fiocruz e outros autores que declaração afiliação Fiocruz na plataforma ORCID) e Lattes (para servidores da Fiocruz que possuem currículo cadastrado). Ao inserir um registro, o pipeline verifica a existência prévia dos autores com base nesses identificadores, assim como verifica a existência prévia de publicações com o mesmo título e ano, para evitar duplicidades. Adicionalmente, os **dados são enriquecidos** com informações relacionadas ao conteúdo das publicações, as áreas de conhecimento do CNPq são vinculadas às produções. Isso é feito de duas formas: i) a primeira é uma rotulação manual, com base na categorização que as unidades da Fiocruz proveram ao Observatório através de planilhas e com as informações coletadas no currículo Lattes dos autores que cadastraram as áreas de conhecimento da publicação; ii) a segunda forma de obtenção desta informação é através da rotulação feita por modelo de Inteligência Artificial (IA). Detalhes deste modelo e deste enriquecimento dos dados serão fornecidos na seção correspondente desta metodologia.

Coleta de dados

Para as bases WoS, Scopus, PubMed, SciELO e Lilacs, a coleta foi realizada diretamente nos respectivos sites utilizando a *string* de busca devidamente adaptada.

Uma *string* de busca (Apêndice I: *String* de busca nas bases de dados) foi elaborada para captar a complexidade e o pluralismo das declarações de afiliação institucional à Fiocruz por parte dos autores em suas produções. A *string* contempla de forma abrangente a heterogeneidade de formas

de escrita da Fiocruz, suas unidades e escritórios regionais, incluindo erros de grafia mais comuns. Como cada base tem sua própria interface, a *string* precisou ser adaptada quanto à sua forma, preservando o mesmo conteúdo, para buscar as ocorrências de publicações nas quais a Fiocruz figurasse entre as afiliações declaradas por seus respectivos autores.

As publicações resultantes dessas buscas foram "baixadas" no formato mais completo disponível em casa base, sendo: WoS e Scopus: *Comma Separated Value* (CSV); PubMed: PubMed Data, Lilacs: Research Information Systems (RIS). Para a SciELO, foi realizado o *download* do arquivo completo da base, onde cada registro é um arquivo *eXtensible Markup Language* (XML). Neste caso, a *string* de busca foi utilizada para encontrar, a partir de um script desenvolvido para tal, a ocorrência da Fiocruz em uma de suas formas em cada arquivo.

Para a coleta na base Arca, que é o repositório Institucional da Fiocruz, não foi necessária consulta por afiliação porque todos os registros ali existentes são produções da Fiocruz. Um conjunto com as publicações do repositório Arca, contemplando o período da análise (jan.2008 jun.2025) foi disponibilizado ao Observatório através de uma parceria com o Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT/Fiocruz).

Por fim, para a base da Plataforma Lattes, foram baixados os currículos em formato XML para cada servidor (ativo e inativo) da Fiocruz. Nesta base foram consultados os currículos Lattes dos servidores a partir do seu CPF, utilizando o 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 de funcionários da instituição. Os dados dos servidores foram fornecidos ao Observatório através de uma parceria interna com a Coordenação Geral de Gestão de Pessoas (COGEPE).

Tratamento e harmonização

Essa é uma das tarefas críticas do processo. Nesta etapa a limpeza, curadoria e harmonização dos dados é realizada. Essa tarefa é feita com auxílio de dicionários de limpeza. Um dicionário é um arquivo que combina chave e valor, onde cada chave tem um valor padronizado. Os dicionários têm sido construídos ao longo dos últimos 5 anos por especialistas do Observatório. São arquivos construídos com auxílio do *software* comercial VantagePoint® e manualmente revisados pela equipe do Observatório. Os dicionários funcionam como arquivos de referência para padronização de diversos campos do banco de dados. Por fim, visam estabelecer um padrão para harmonização e permitir a soma mais precisa dos quantitativos de igual valor. No exemplo a seguir *key* representa a forma como o autor declarou sua afiliação, e *value* a forma como foi normalizado o nome da unidade.

key	value
Oswaldo Cruz Fdn FIOCRUZ, Ctr Technol Dev Hlth CDTS, Natl Inst Sci & Technol Innovat Neglected Populat, Av Brasil 4365, BR-21040900 Rio De Janeiro, RJ, Brazil	Fiocruz/Centro de Desenvolvimento Tecnológico em Saúde
Oswaldo Cruz Fdn FIOCRUZ, Oswaldo Cruz Inst, Lab Epidemiol & Mol Systemat LESM, BR-21040900 Rio De Janeiro, Brazil	Fiocruz/Instituto Oswaldo Cruz
Oswaldo Cruz Fdn FIOCRUZ BA, Goncalo Moniz Inst IGM, Ctr Data & Knowledge Integrat Hlth CIDACS, Salvador, BA, Brazil	Fiocruz/Instituto Gonçalo Moniz
Fiocruz MS, Oswaldo Cruz Inst, Lab Comparat & Environm Virol, BR-21040360 Rio De Janeiro, RJ, Brazil	Fiocruz/Instituto Oswaldo Cruz
Fiocruz. Escola Nacional de Saúde Pública. Centro Latino-Americano de Estudos de Violência e Saúde Jorge Carelli. Rio de Janeiro. BR	Fiocruz/Escola Nacional de Saúde Pública Sérgio Arouca
Fiocruz MS, Far Manguinhos, Av Comandante Guaranys 447, BR-22775903 Rio De Janeiro, RJ, Brazil	Fiocruz/Instituto de Tecnologia em Fármacos (Farmanguinhos)
Fiocruz MS, LapClin DST AIDS, Natl Inst Infectol, INI, BR-21045900 Rio De Janeiro, Brazil	Fiocruz/Instituto Nacional de Infectologia Evandro Chagas
Fiocruz MS, Mycol Lab, Evandro Chagas Clin Res Inst, BR-21045900 Rio De Janeiro, Brazil	Fiocruz/Instituto Nacional de Infectologia Evandro Chagas

Para garantir a melhor curadoria dos dados foram criados dicionários para os campos: Unidade Fiocruz, Paceiros, Tipologia documental, Veículos, Autores, Países e ISSN, sendo este último construído como auxílio da base OpenAlex.

O dicionário de veículos de publicação merece destaque. Muitas produções científicas são oriundas de comunicações em simpósios, congressos e eventos. Este conteúdo, quando coletado da Plataforma Lattes, embora riquíssimo, traz associado uma série de ambiguidades devido ao livre preenchimento dos dados nesta base. Adotou-se então, um critério de junção dos eventos acadêmicos independentemente da sua localidade ou de seu ano de ocorrência no dicionário de veículos de publicação. Por exemplo, o Congresso da Sociedade Brasileira de Parasitologia ocorreu em 2009 com o nome XXI Congresso da Sociedade Brasileira de Parasitologia, em 2011 XXII Congresso da Sociedade Brasileira de Parasitologia, em 2013 como XXIII Congresso da Sociedade

Brasileira de Parasitologia. Estas variações, além das variações não oficiais, abreviações e erros de digitação preenchidas pelos autores, tais como 22o Congresso de Parasitologia, foram harmonizados sobre o nome guarda-chuva: Congresso da Sociedade Brasileira de Parasitologia, pois essa informação é exibida junto com o ano de publicação, então não há prejuízo quanto a identificação da edição do evento.

Além dos dicionários, outros dados de referência também são utilizados no formato de planilhas. É o caso de dados de pessoal, fornecidos pela COGEPE, área de Recursos Humanos da Fiocruz e a tabela de áreas do conhecimento do CNPq, na qual cada publicação foi classificada pelas unidades da Fiocruz em uma planilha e imputada posteriormente no banco de dados. Este trabalho de identificação das áreas é recente, oriundo de uma parceria com a Vice-Presidência de Pesquisa e Coleções Biológicas (VPPCB).

Por fim, é apresentada no Apêndice II: Sistematização das palavras-chaves do autor, uma lista de palavras-chaves dos autores que foram consideradas sinônimos para melhor descrever a temática de concentração dos trabalhos publicados.

Identificação das unidades Fiocruz e seus parceiros

Para construir os dicionários de afiliações (27 unidades e/ou escritórios, além de parceiros), inicialmente foi aplicado um modelo de aprendizado de máquina treinado para reconhecer se uma afiliação é ou não Fiocruz. Para o grupo reconhecido como Fiocruz foi construído o dicionário de unidades da Fiocruz. Para o grupo reconhecido como não Fiocruz foi construído o dicionário de parceiros. Ambos os casos com curadoria manual e semiautomatizada via *software* comercial Vantage Point®. Este processo se aplica às seis primeiras bases analisadas (exceção a Plataforma Lattes,) aplicando os dicionários construídos a partir das instituições que cada autor declarou na publicação como afiliações.

Para a sétima base, a Plataforma Lattes, a afiliação considerada foi aquela assinalada nos dados recebidos da COGEPE. Esta diferença de procedimento justifica-se porque os dados do Lattes apresentam inconsistências quanto a afiliação dos autores, possivelmente derivadas de sua característica de autopreenchimento. Por exemplo, muitos autores preenchem apenas Fiocruz como afiliação em seus currículos Lattes, sem indicar ou "dar pistas" da unidade a qual pertencem. Esta forma de vincular o servidor a sua produção é estática não levando em conta nesta para os dados desta base, as migrações entre unidades, por exemplo. Contudo, é a forma mais efetiva de se captar a afiliação em nível de unidade Fiocruz, quando a outra opção é a apenas Fiocruz. Este procedimento de se afirmar que estas publicações são institucionais só foi possível uma vez que coletamos apenas os currículos de servidores, que por tanto, asseguram sua afiliação com a Fiocruz.

É importante mencionar que há um quantitativo grande de produções científicas que ainda estão agrupadas com o rótulo de unidade "Fiocruz". Este número se refere a produções para as quais não foram identificadas, em um primeiro tratamento, a unidade a qual pertence o autor. A dificuldade de padronização de nomes institucionais é um problema comum, ainda sem solução totalmente eficaz mesmo com uso de modernas técnicas bibliométricas e computacionais. Desse modo, é de extrema relevância que se padronize a forma de se citar a instituição e a unidade ao qual os autores

tem vínculo. Assim será possível melhorar a identificação das publicações de cada unidade da Fiocruz.

Linhas gerais dos processos de harmonização

Limpeza e Processamento de Dados

- Validação de títulos: Verificação se o título é válido e possui tamanho adequado
- Validação de ano: Confirmação se o ano de publicação está dentro do intervalo desejado (2008-2025)
- Tratamento de valores nulos: Conversão de valores NaN para strings vazias

Limpeza de Texto

O sistema aplica múltiplas camadas de limpeza:

- Correção de Caracteres
    - Substituição de caracteres mal codificados (ex: "Ã£" → "ã")
    - Remoção de caracteres especiais indesejados
    - Harmonização de codificação UTF-8

Padronização de Strings

- Remoção de aspas no início e fim das strings
- Eliminação de tags HTML
- Remoção de pontuação indesejada
- Padronização de espaços múltiplos
- Remoção de acentos quando necessário

Tratamento Específico

- Remoção de números ordinais
- Eliminação de numerais romanos
- Remoção de formatos de data
- Tratamento de palavras com apenas consoantes
- Remoção de partes numéricas em eventos científicos

Geração de Identificadores Únicos

- Hash Fiocruz: Geração de identificador único baseado no título e ano de publicação
- Hash de afiliação: Identificador único para cada afiliação processada
- Hash de autor: Identificador único para cada autor

Dicionários de Referência

O sistema utiliza múltiplos dicionários auxiliares para harmonização:

- Unidades Fiocruz: Mapeamento de variações de nomes das unidades Fiocruz
- Parceiros: Dicionário de instituições parceiras
- Tipos de documento: Classificação padronizada de tipos de publicação
- ISSN: Mapeamento de ISSNs para títulos de periódicos
- Veículos: Harmonização de nomes de periódicos
- Países: Códigos de países padronizados

Harmonização de Afiliações Fiocruz

- Afiliações Fiocruz: Aplicação de dicionário de unidades Fiocruz com correspondência exata e fuzzy matching
- Verificação se a afiliação pertence à Fiocruz através de lista extensa de variações
- Aplicação de critérios para evitar falsos positivos (ex: escolas em Portugal)

Harmonização de Afiliações Parceiras

- Afiliações Parceiras: Harmonização usando dicionário de parceiros com correspondência exata e fuzzy matching
- Melhoria com IA: Uso de API Maritaca.ai para melhorar nomes de afiliações
- Enriquecimento ROR: Consulta à API ROR.org para obter informações adicionais

Processamento de Autores

- Separação de autores e suas respectivas afiliações
- Formatação padronizada de nomes (inversão de ordem quando necessário)
- Associação de ORCIDs e Researcher IDs
- Geração de identificadores únicos

Harmonização de Veículos de Publicação

- Correspondência por ISSN (impresso e eletrônico)
- Harmonização de títulos de periódicos
- Aplicação de critérios de qualidade

Processamento de Palavras-chave

- Separação de palavras-chave por ponto e vírgula
- Remoção de duplicatas
- Capitalização adequada
- Classificação por tipo (palavras-chave do autor vs. Keywords Plus)

Conversão de formato

Cada base, a partir dos distintos formatos coletados, fornece seus dados organizados em diferentes campos, que por sua vez, abrigam dados sob diferentes arranjos. Para solucionar este problema, foram desenvolvidos *scripts in-house* para ler cada registro de cada base e convertê-los num novo arquivo em formato JSON, homogeneizando os campos e suas formatações. A escolha dos campos e seu arranjo em um novo arquivo JSON foram adaptadas da especificação utilizada pela base DOA²J.

Quando cada registro é convertido para o formato intermediário JSON, alguns tratamentos iniciais são aplicados para mitigar problemas como acentuação por uso de diferentes codificações de caracteres, presença de *tags* de formatação nos dados (html, XML, URL, formatações de data, por exemplo). Adicionalmente, são aplicados filtros para eliminar registros incompletos, tais como, sem ano, sem autores e sem títulos. Cada arquivo convertido é salvo com seu fiocruz_id, que é um nome representando um identificador único, gerado para cada publicação a partir de uma função hash que combina o título e o ano de publicação, tal como vieram das bases, em uma *string* única com tamanho fixo de 40 caracteres, por exemplo, 5802e1cd018410762da65ef52892c7e0931e5627.json. A criação do fiocruz_id é de extrema importância, pois este é um identificador único persistente, que permite determinar com menos recursos computacionais duplicidades em publicações e fornecer um meio confiável de indexação.

Limpeza dos dados

Importante salientar que, enquanto as demais bases são orientadas à publicação, a Plataforma Lattes é orientada a autor. Esta diferença, representa um passo extra da coleta e tratamento inicial de dados para combinar uma mesma publicação declarada em diferentes currículos Lattes antes da conversão para o formato JSON. Ainda, como há currículos Lattes de servidores ativos e inativos foram consideradas as datas de entrada e eventual saída no caso de inativos – dos servidores de forma a compatibilizar a produção com o período delimitado. Ao converter os registros de cada base, alguns **critérios de exclusão são aplicados**: i) Registros sem títulos ou títulos com menos de 10 caracteres; ii) Registros sem ano de publicação ou com ano de publicação diferente do intervalo 2008-2025; iii) Registros sem autor.

Dessa forma, após cada registro de cada base passar pelos filtros de exclusão e ser convertido para o formato JSON com seu respectivo fiocruz_id, o conjunto de dados resultando desta etapa passa a ser estruturado e atende aos requisitos para inserção no banco de dados.

A principal forma de limpeza dos dados é aplicação de dicionários por meio *de scripts in-house* desenvolvidos para tal tarefa. Nesta etapa cada JSON "sujo" é submetido a uma limpeza, com critérios de classificação e de exclusão, gerando um JSON "limpo", que está pronto para ser inserido no banco de dados.

Ao processar cada registro, são aplicados os dicionários de afiliações (unidades da Fiocruz e instituições parceiras), de veículos de publicação (ISSN e nomes de veículos) e tipos de documentos. Vale ressaltar que quanto a tipologia documental, são consideradas as mais diversas tipologias: artigos, congressos, resumo, eventos, capítulo de livro, livros dentre outras. As estratégias de comparação do dado de um registro com os dados dos dicionários são: busca exata, busca aproximada utilizando *lógica fuzzy*[3], busca aproximada por distância de Levenshtein[4]. Com esta estratégia conseguimos padronizar os dados estabelecendo uma convergência da diversidade de nomes fornecidos.

Após a aplicação dos dicionários, diversos registros são descartados e não serão criados arquivos JSON limpos para estes registros. Os critérios de remoção nesta etapa são:

- Tipologia documental: errata, carta ao autor, *corrected, republished article, preprint, address, comment, correction, erratum, corrigendum retracted* foram desconsiderados do conjunto final de dados. Estes não são considerados para o *dataset* final, na intenção de favorecer as publicações Institucionais inéditas.
- Publicações sem autores vinculados à Fiocruz: uma vez aplicados os dicionários, é possível identificar com mais precisão qual é a afiliação de cada autor que fora declarada na publicação. Se não houver entre as afiliações declaradas de cada autor, pelo menos uma que seja Fiocruz, o registro é descartado. Este critério corrige muitos problemas de falsos positivos advindos das bases durante a coleta, por exemplo "Escola Nacional de Saúde Pública", localizada em Portugal e pertencente a Universidade Nova de Lisboa, "Instituto Evandro Chagas" situado no Pará, e o "Hospital Oswaldo Cruz", hospital universitário da Universidade de Pernambuco instalado no Recife. Estes exemplos, poderiam facilmente serem confundidos, com as unidades da Fiocruz: ENSP; INI e Aggeu Magalhães respectivamente.

- Nomes de autores: especificamente para o caso de autores, também foi desenvolvido um script que normaliza a forma de apresentação dos nomes, levando em conta as variações em citações, existência de homônimos e abreviações. Tomando uma pessoa chamada “Beltrano Cicrano da Silva”, se o nome ocorrer desta forma, assim permanecerá. Se o nome ocorrer abreviado como: “B C Silva”, ou “Silva, BC”, ou “Silva B. C.”, será padronizado para “B. C. Silva”. Este script homogeneíza essas variações, para a forma onde é possível considerando que há nomes homônimos.

Armazenamento dos dados

Um banco de dados foi especialmente desenhado para abrigar estes dados, tendo como sistema gerenciador de bancos de dados (SGBD), o PostgreSQL4, um banco de dados relacional *Open Source* e bem estabelecido no mercado. Foram construídos scripts que mapeiam os dados para o banco de dados através de uma biblioteca Python chamada sqlalchemy para Object Relational Model (ORM).

As primeiras informações inseridas são os autores oriundos da COGEP, com suas respectivas unidades de afiliação e número de ORCID. Os demais autores também são buscados via API[5] ORCID. Esta inserção preliminar visa evitar duplicidades de autores, a partir da checagem prévia no banco de dados. O passo seguinte foi o armazenamento dos dados das publicações propriamente ditas oriundos das coletas das bases, a partir de seus JSON limpos. A inserção no banco de dados obedece a uma ordem de precedência. Essa ordem de inserção preferencial foi estabelecida em função da qualidade e completude dos campos dos registros coletados a saber: 1ª WoS; 2ª Scopus; 3ª PubMed; 4ª SciELO; 5ª Lilacs; 6ªArca e por último o 7ª Plataforma Lattes. Ao inserir um registro oriundo de uma base específica, o sistema verifica se a publicação já está presente no banco de dados, a fim de evitar inserções duplicadas. Os critérios para checagem de publicações repetidas são os seguintes:

- Idêntico fiocruz_id,
- Idêntico DOI (caso disponível), (a completude deste campo gira em torno de 75%).
- Idêntico título e ano (título convertido para letras minúsculas e sem acentuação),
- Idêntico título e idêntico veículo de publicação (título convertido para letras minúsculas e sem acentuação).

Se o registro ainda não está presente no banco de dados, ele é inserido assinalando a base de origem e aproveitando os autores preexistentes no banco de dados, caso seja possível. Se o registro já está presente, ele não é inserido, e apenas a informação da base de origem é atualizada, acrescendo a nova origem.

Dados das áreas de conhecimento de acordo com o CNPq

O objetivo da categorização de cada publicação, por áreas de conhecimento, é identificar e vincular à produção científica da Fiocruz, as áreas de pesquisa que a instituição se dedica e tem mais vocação. A classificação padronizada para identificação das áreas de conhecimento pautou-se pela adoção da classificação hierárquica de áreas do CNPq, considerando que esta é consolidada, amplamente conhecida e utilizada por todos os pesquisadores do país. Essa informação foi obtida de duas formas diferentes:

Diretamente das unidades ou da Plataforma Lattes: Representantes das unidades da Fiocruz categorizaram uma amostra das suas publicações com a tipologia documental artigo. Também foram consideradas as classificações realizadas diretamente na Plataforma Lattes. Essa categorização foi vinculada as áreas de conhecimento do CNPq. O sistema possui quatro níveis hierárquicos que permitem uma análise detalhada da área de conhecimento de cada publicação. Esses níveis são: Grande área, Área, Subárea e Especialidade, também chamados de nível 1,2,3 e 4, respectivamente.

Uso de IA: adicionalmente, um modelo de Inteligência Artificial foi desenvolvido para auxiliar na categorização dos dados. Este modelo de IA rotulou as produções automaticamente. Os detalhes sobre o modelo de IA estão no artigo "*Hierarchical Article Classification: A Multi-Level Framework for Organizing Scholarly Literature*", disponível sob o DOI: 10.1109/ACCESS.2025.3579232.

Para ambas as formas de obtenção da rotulação do CNPq (Manual ou automatizada), foi necessária uma revisão para adequação das áreas dentro da hierarquia definida pelo CNPq. Essa adaptação foi feita por meio de *scripts*, em Phyton3, desenvolvidos *in-house*. Os ajustes foram feitos no sentido de corrigir erros de indentação dos níveis. Por exemplo, algumas publicações foram categorizadas com Grande área Antropologia. No entanto, de acordo com o CNPq, a Antropologia é uma Área dentro da Grande área de Ciências Humanas. Logo, foram realizadas correções. Vale ressaltar que as unidades categorizam algumas publicações com "novas áreas", sugerindo, portanto, a necessidade da ampliação e revisão desta categorização proposta pelo CNPq. Essas publicações não foram consideradas a *priori* no dashboard. Como exemplo, pode-se citar: Bioinformática, Tecnologias Assistidas, Direito Sanitário, dentre outras sugestões. Vale ressaltar que a informação advinda das definições das unidades e da Plataforma Lattes representa algo próximo a 55% do total das produções.

Enriquecimento dos dados

Para cada produção, cujo veículo de publicação tinha ISSN cadastrado no DOAJ, foram importados para o banco de dados os metadados correspondentes. Os dados de ORCID (DOI: 10.21680/2447-7842.2023v9n2ID33661), para autores, e ROR, para instituições, foram vinculados a cada registro do banco de dados sempre que o vínculo pôde ser estabelecido. A base Open Alex foi usada para harmonizar os nomes de veículos de publicação.

Dashboard

A partir dos dados armazenados em banco de dados próprio, foi possível, utilizando a tecnologia Microsoft Power BI, construir o painel interativo que esta metodologia acompanha.

Curadoria

Os dados do banco de dados passam semanalmente por revisão manual, de forma que, incrementalmente são acrescentadas melhorias advindas de correções ou complementação de dados. Tais melhorias são refletidas no *dashboard* que está programador para atualizar-se a partir do banco de dados, também com frequência semanal.

Apêndice I: *String* de busca nas bases de dados

("Biblioteca Manguinhos" OR "Bio Manguinhos" OR "Biomanguinho" OR "Biomanguinhos" OR "Inst Tecnol Imunobiol Biomanguinhos" OR "Inst Tecnol Imunobiol/Fiocruz" OR "Inst Tecnol Imunobiol-Fiocruz" OR "Bio Manguinhos" OR "Biomanguinhos" OR "Bio-Manguinhos" OR "Biomanguinlios" OR "Inst Tecnol Imunobiol" OR "Instituto de Tecnologia em Imunobiológicos" 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 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 "Ctr Desenvolvimento Tecnol Saude" 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 Osvaldo Cruz" OR "Casa de Osvaldo 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 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 Osvaldo Crus" OR "Fundação Oswaldo Cruz" OR "Fundação Osaldo Cruz" OR "Fundação Osawaldo Cruz" OR "Fundação Osqaldo Cruz" OR "Fundação Osvaldo Cruz" OR "Fundação Oswald Cruz" OR "Fundação Oswaldo Crus" OR "Fundação Oswaldo Crusz" OR "Fundação Oswaldo Curz" OR "Fundação OSWALDO CURZ" OR "Fundação Oswalo Cruz" OR "Fundação Owaldo Cruz" OR "Fdn Osvaldo Crus" OR "Fdn Oswaldo Cruz" OR "Fdn Osaldo Cruz" OR "Fdn Osawaldo Cruz" OR "Fdn Osqaldo Cruz" OR "Fdn Osvaldo Cruz" OR "Fdn Oswald Cruz" OR "Fdn Oswaldo Crus" OR "Fdn Oswaldo Crusz" 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 "FlOCRUZ" OR "Fiocruz" OR "Osvaldo Crus Foundation" OR "Oswaldo Cruz Foundation" OR "Osaldo Cruz Foundation" OR "Osawaldo Cruz Foundation" OR "Osqaldo Cruz Foundation" OR "Osvaldo Cruz Foundation" OR "Oswald Cruz Foundation" OR "Oswaldo Crus Foundation" OR "Oswaldo Crusz 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 "Fiocruz/Amazonia" OR "FiocruzAmazonia" OR ILMD OR "Fiocruz-Manaus" OR "Fiocruz/Manaus" OR "CPGM" OR "Goncalo Moniz" OR "Goncalo Muniz" OR “Instituto Gonçalo Moniz” OR “Gonçalo Moniz”OR "Fiocruz/Bahia" OR "FiocruzBahia" OR "Fiocruz/BA" OR "Fiocruz-BA" OR "Fiocruz/Salvador" OR "Fiocruz-Salvador" OR "FiocruzBrasília" OR "Fiocruz-DF" OR "Fiocruz/Brasília" OR "Fiocruz/DF" OR "DIREB" OR "GEREB" OR "Gerência Regional de Brasília" OR "Fiocruz/Ceará" OR "Fiocruz-Ceará" OR "Fiocruz/Fortaleza" OR "FiocruzFortaleza" OR "Fiocruz/CE" OR "Fiocruz-CE" OR "Fiocruz/Mato Grosso" OR "Fiocruz-Mato Grosso" OR "Fiocruz/MS" OR "Fiocruz-MS" OR "Fiocruz/Campo Grande" OR "Fiocruz-Campo Grande" OR "Rene Rachou" OR "Renne Rachou " OR "Rene Rachu" OR "Renne Rachu " OR "Cpqrr" OR "Fiocruz/MG" OR "FiocruzMG" OR "Fiocruz/Minas" OR "Fiocruz-Minas" OR "Fiocruz/Belo Horizonte" OR "Fiocruz-Belo Horizonte" OR "Fiocruz/BH" OR "Fiocruz-BH" OR “Instituto René Rachou” OR "Ageu Magalhaes" OR "Aggeu Magalhaes" OR "Cpqam" OR "Fiocruz/Pernambuco" OR "Fiocruz-Pernambuco" OR "Fiocruz/PE" OR "FiocruzPE" OR "Fiocruz/Recife" OR "Fiocruz-Recife" OR "Fiocruz/Piaui" OR "Fiocruz-Piaui" OR "Fiocruz/PI" OR "FiocruzPI" OR "Fiocruz/Teresina" OR "Fiocruz-Teresina" OR "Instituto Carlos Chagas/Fiocruz" OR "Instituto Carlos ChagasFiocruz" OR "Carlos Chagas Inst/Fiocruz" OR "Carlos Chagas Inst-Fiocruz" OR "Carlos Chagas Institute/Fiocruz" OR "Carlos Chagas Institute-Fiocruz" OR "ICC-Paraná" OR "ICC/Paraná" OR "ICC-PR" OR "ICC/PR" OR "ICCCuritiba" OR "ICC/Curitiba" OR "FIOCRUZ-Paraná" OR "FIOCRUZ/Paraná" OR "FIOCRUZPR" OR "FIOCRUZ/PR" OR "FIOCRUZ-Curitiba" OR "FIO-CRUZ/Curitiba" OR "ICCFiocruz" OR "ICC/Fiocruz" OR "Fiocruz/Rondonia" OR "Fiocruz-Rondonia" OR "Fiocruz/RO" OR "FiocruzRO" OR "Fiocruz/Porto Velho" OR "Fiocruz-Porto Velho" OR "Instituto de Comunicação e Informação Científica e Tecnológica em Saúde" OR "ICICT" OR "CICT" OR "Institute of Communication and Scientific and Technological Information in Health" OR "Instituto de Ciência e Tecnologia em Biomodelos" OR "Institute of Science and Technology in Biomodels" OR "Centro de Criação de Animais de Laboratório" OR "CECAL" OR "ICTB" OR "Fernandes Figueira" OR "Fernandes Fugueira" OR "Fernandes Figueiras" OR "Fernandez Figueira" OR "IFF/Fiocruz" OR "IFFFiocruz" OR "Instituto Nacional de Saúde da Mulher, da Criança" OR "National Institute of Health for Women, Children" OR "Instituto de Saúde da Mulher, da Criança" OR "Institute of Health for Women, Children/Fiocruz" OR "Institute of Health for Women, Children-Fiocruz" OR "Instituto Nacional de Controle de Qualidade em Saúde" OR "Instituto de Controle de Qualidade em Saúde" OR "National Institute for Quality Control in Health" OR "Inst Nacl Controle Qualidade Saude" OR "INCQS" OR "Evandro Chagas Inst/Fiocruz" OR "Evandro Chagas Inst-Fiocruz" OR "Evandro Chagas/Fiocruz" OR "Evandro Chagas-Fiocruz" OR "Ipec/Fiocruz" OR "Ipec-Fiocruz" OR "Instituto Nacional de Infectologia" OR "National Institute of Infectious Diseases/Fiocruz" OR "National Institute of Infectious DiseasesFiocruz" OR "Instituto de Pesquisa Clínica Evandro Chagas" OR "Evandro Chagas Clinical Research Institute" OR "INI/Fiocruz" OR "INI-Fiocruz" OR "Inst Pesquisa Clin Evabdro Chagas" OR "Inst Pesquisa Clin Evandro Chagas" OR "Inst Oswaldo Fdn" OR "IOC/Fiocruz" OR "IOC-Fiocruz" OR "Instituto Osvaldo Crus" OR "Instituto Oswaldo Cruz" OR "Instituto Osaldo Cruz" OR "Instituto Osawaldo Cruz" OR "Instituto Osqaldo Cruz" OR "Instituto Osvaldo Cruz" OR "Instituto Oswald Cruz" OR "Instituto Oswaldo Crus" OR "Instituto Oswaldo Crusz" OR "Instituto Oswaldo Curz" OR "Instituto OSWALDO CURZ" OR "Instituto Oswalo Cruz" OR

"Instituto Owaldo Cruz" OR "Inst Osvaldo Crus" OR "Inst Oswaldo Cruz" OR "Inst Osaldo Cruz" OR "Inst Osawaldo Cruz" OR "Inst Osqaldo Cruz" OR "Inst Osvaldo Cruz" OR "Inst Oswald Cruz" OR "Inst Oswaldo Crus" OR "Inst Oswaldo Crusz" OR "Inst Oswaldo Curz" OR "Inst OSWALDO CURZ" OR "Inst Oswalo Cruz" OR "Inst Owaldo Cruz" OR "Osvaldo Crus Institute" OR "Oswaldo Cruz Institute" OR "Osaldo Cruz Institute" OR "Osawaldo Cruz Institute" OR "Osqaldo Cruz Institute" OR "Osvaldo Cruz Institute" OR "Oswald Cruz Institute" OR "Oswaldo Crus Institute" OR "Oswaldo Crusz Institute" OR "Oswaldo Curz Institute" OR "OSWALDO CURZ Institute" OR "Oswalo Cruz Institute" OR "Owaldo Cruz Institute" OR "Osvaldo Crus Inst" OR "Oswaldo Cruz Inst" OR "Osaldo Cruz Inst" OR "Osawaldo Cruz Inst" OR "Osqaldo Cruz Inst" OR "Osvaldo Cruz Inst" OR "Oswald Cruz Inst" OR "Oswaldo Crus Inst" OR "Oswaldo Crusz Inst" OR "Oswaldo Curz Inst" OR "OSWALDO CURZ Inst" OR "Oswalo Cruz Inst" OR "Owaldo Cruz Inst" OR "Procc/Fiocruz" OR "Procc-Fiocruz" OR "Programa Comp Cient Qswald Cruz")

Apêndice II: Sistematização das palavras-chaves do autor

Para uma melhor sistematização das palavras-chave citadas pelos autores em suas publicações, as seguintes palavras foram agrupadas.

Key	Value
Adolescente	Adolescente
Adolescent	Adolescente
Amazônia	Amazônia
Amazon	Amazônia
Atenção Primária à Saúde	Atenção primária à saúde
Atenção primária à saúde	Atenção primária à saúde
Brazil	Brasil
Brasil	Brasil
COVID-19	Covid-19
Covid-19	Covid-19
Children	Criança
Child	Criança
Diagnóstico	Diagnóstico
Diagnosis	Diagnóstico
Epidemiology	Epidemiologia
Epidemiologia	Epidemiologia
Inflammation	Inflamação
Inflamação	Inflamação
Malária	Malária
Malaria	Malária
Mortality	Mortalidade
Mortalidade	Mortalidade
Prevalência	Prevalência

Prevalence	Prevalência
Primary Health Care	Primary health care
Primary health care	Primary health care
Saúde do trabalhador	Saúde do trabalhador
Saúde do Trabalhador	Saúde do trabalhador
Saúde pública	Saúde pública
Saúde Pública	Saúde pública
Public health	Saúde pública
Unified Health System	Sistema Único de Saúde
SUS	Sistema Único de Saúde
Sistema Único de Saúde	Sistema Único de Saúde
Tuberculosis	Tuberculose
Tuberculose	Tuberculose
Violência	Violência
Violence	Violência
Zika virus	Zika
Zika	Zika