METODOLOGIA
COLETA E TRATAMENTO DOS DADOS DE PRODUÇÃO CIENTÍFICA DA FUNDAÇÃO OSWALDO CRUZ

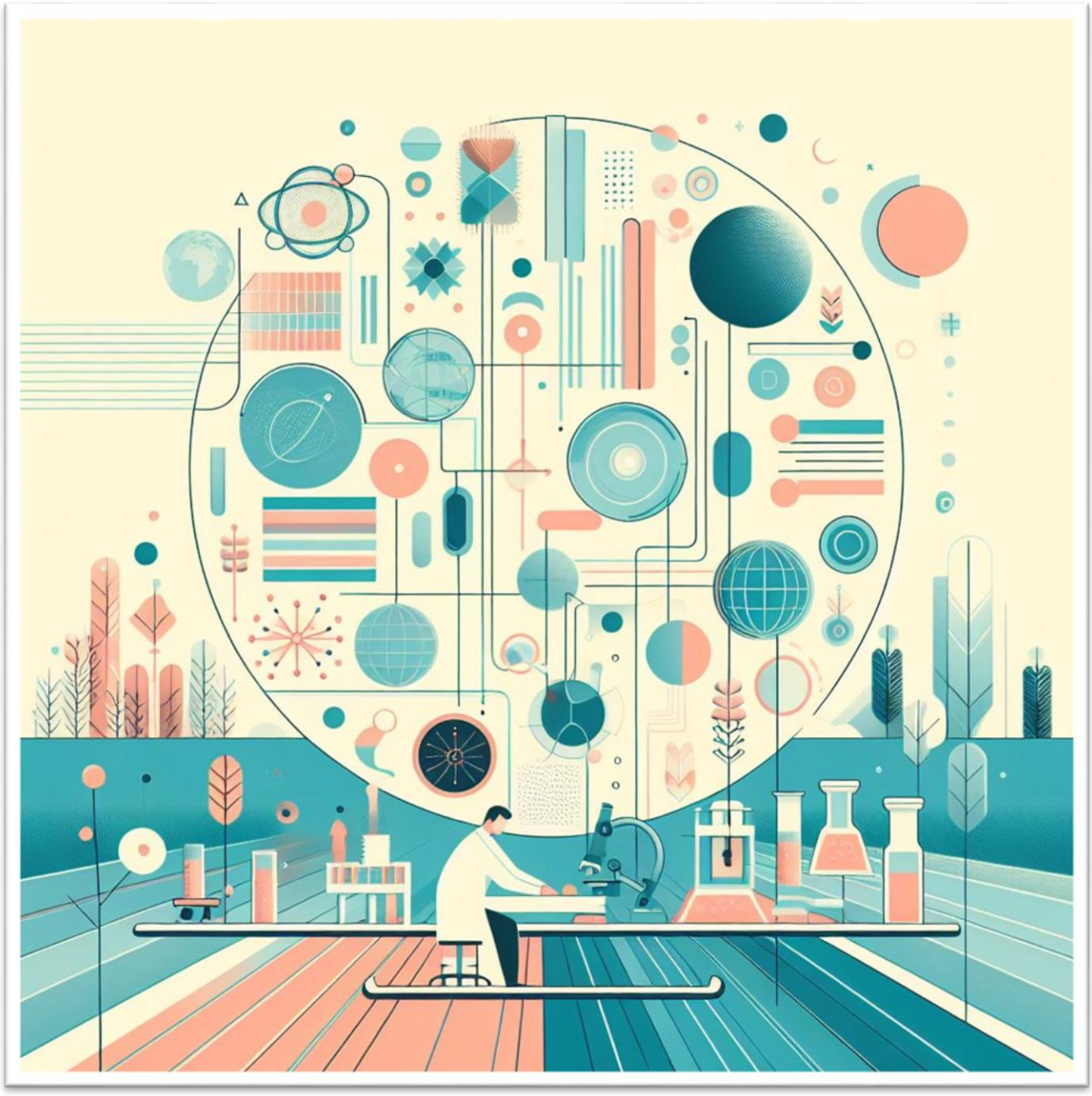


Figura gerada com IA · 17 de março de 2024 às 8h38 utilizando *Microsoft Copilot Designer*

JAN.2008 A DEZ.2023

SUMÁRIO

INTRODUÇÃO

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 indiretamente as ações que compõe o Sistema Único de Saúde.

A comunicação gráfica é uma maneira eficaz de se comunicar resultados. *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. Este documento descreve os procedimentos adotados na coleta, transformação e disponibilização dos indicadores de produção científica da Fiocruz em um *dashboard*. Seu objetivo é promover a clareza, 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 dezembro de 2023. Este indicador é, portanto, um compilado dos últimos 17 anos de produção científica de uma das mais renomadas e respeitadas instituições públicas do Brasil: a Fundação Oswaldo Cruz.

VISÃO GERAL DA METODOLOGIA

Os dados de publicações técnicas e científicas foram **coletados**, **tratados**, **enriquecidos** e **inseridos em um banco de dados** a partir do qual foi construído um *dashboard* com indicadores de produção técnico-científica. A **Erro! Fonte de referência não encontrada.** apresenta uma visão geral do *pipeline* que é descrito em detalhes nesta metodologia.

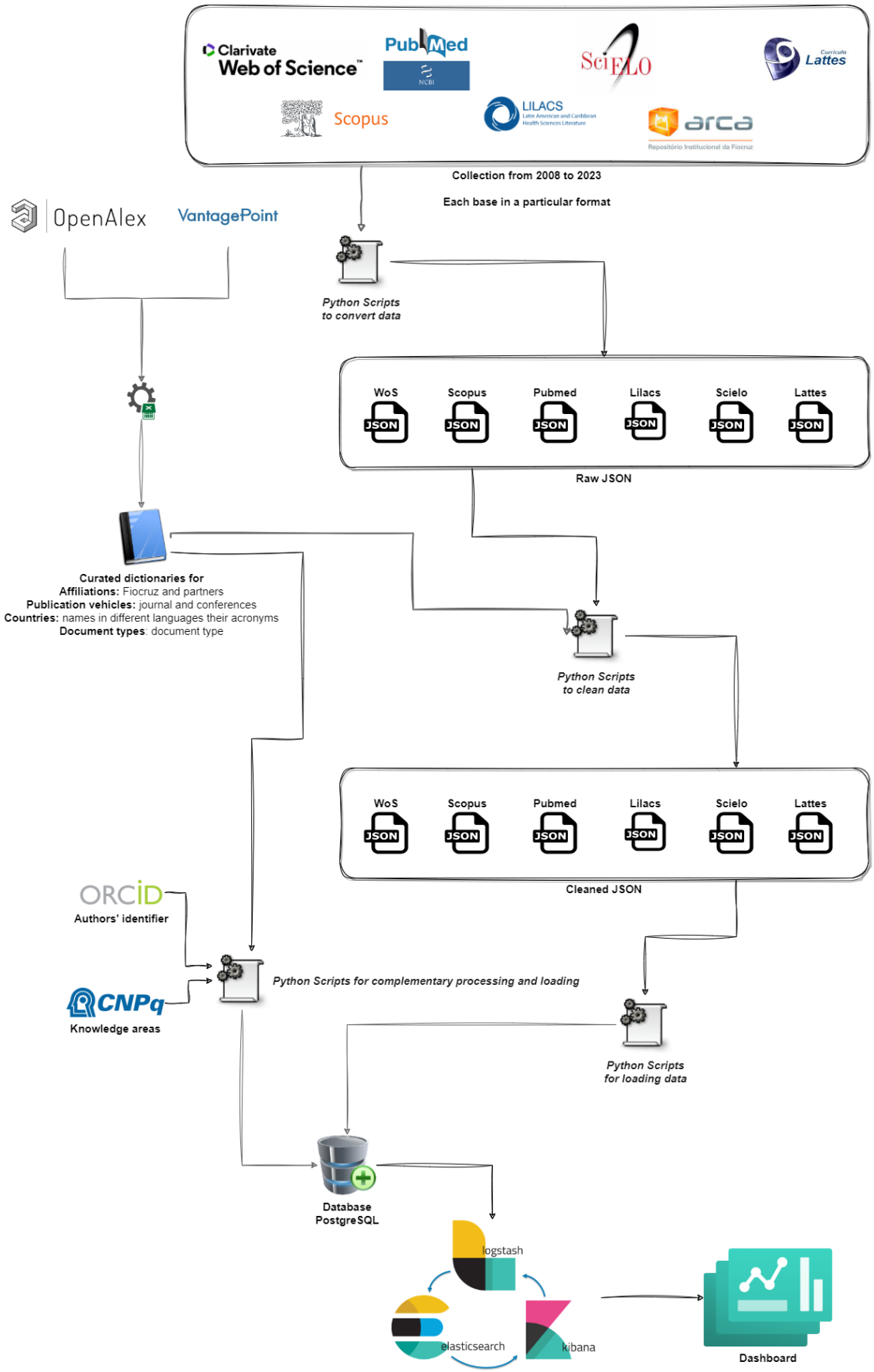
A coleta ocorreu em sete bases de dados: Web of Science (WoS), Scopus, PubMed, Lilacs, SciELO, Arca e Currículo Lattes. Nas bases WoS, Scopus, PubMed e Lilacs, foi utilizada uma *string* de busca por afiliação, comum a todas as bases, apenas com adaptação no formato. A base SciELO foi completamente baixada em arquivos no formato XML, aplicando a cada arquivo a mesma *string* de busca de afiliação. Os dados do Arca foram e coletados pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) em formato Excel e disponibilizados para análise. Da base Currículo Lattes foram extraídas informações dos currículos dos servidores ativos e inativos da Fiocruz com base em seus CPF. Adicionalmente, foram coletados dados complementares de bases de apoio, a saber: ORCID, OpenAlex, DOAJ e CrossRef, esta etapa tem a função de auxiliar no estabelecimento de vínculos entre os dados e melhorar a completude da informação coleta.

O tratamento dos dados envolveu diversas etapas organizadas em um *pipeline*. Cada coleta tem formatos e estruturas de dados distintos, 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 foram do intervalo pré-estabelecido 2008-2023, publicações do Lattes que respeitem o intervalo de permanência dos servidores na Fiocruz, exclusão de 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 capazes de padronizar alguns dados, identificar irregularidades e corrigi-las ou sinalizá-las para correção manual ou com auxílio do *software* 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 com base na categorização que as unidades da Fiocruz proveram ao Observatório.



A seguir detalhes de cada uma das fases são apresentados para o melhor entendimento do processo.

COLETA DE DADOS

Uma *string* de busca (Anexo 1) foi especialmente construída 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. A string foi utilizada em 5 das 7 bases bibliográficas: WoS, Scopus, PubMed, SciELO e Lilacs. 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.

Para as bases *WoS, Scopus, PubMed, SciELO* e *Lilacs*, a coleta foi realizada diretamente nos respectivos sites utilizando a *string* de busca devidamente adaptada. As publicações resultantes dessas buscas foram "baixadas" no formato mais completo disponível em casa base, sendo: WoS: Comma Separated Value (CSV); Scopus: Comma Separated Value (CSV); PubMed: PubMed Data[1], 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 dez.2023) 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 a 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 DOS DADOS

No contexto do Observatório, um dos recursos mais importantes no tratamento dos dados são os dicionários. Para o Observatório, um dicionário é um arquivo que combina chave e valor, onde cada chave tem um valor recomendado. Os dicionários têm sido construídos ao longo dos últimos 3 anos por especialistas do Observatório e funcionam como arquivos de referência para diversas situações como nomes de veículos de publicação, afiliações, tipos de documentos, entre outros. Dessa forma, os

[1] https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ad

dicionários configuram-se como conjuntos de dados manualmente curados, que visam estabelecer uma *baseline* de conhecimento sobre um tema em específico. Exemplo: quais são as variações com que uma unidade da Fiocruz é indicada como afiliação em trabalhos científicos. Em geral, são arquivos binários do tipo *pickle* construídos a partir de planilhas e arquivos *thesaurus* (.the) obtidos a partir de agrupamentos realizados no software VantagePoint[2]. Os dicionários disponíveis são:

- **aux_fiocruz_units**: 184.681 registros catalogados de formas distintas de preenchimento de afiliação para alguma unidade da Fiocruz.
- **aux_partners**: 147.550 registros catalogados de formas distintas de preenchimento de afiliação para alguma instituição parceira em publicações com pelo menos um autor da Fiocruz.
- **aux_document_type**: 198 tipos catalogados de documentos, tais como Artigo, Article, Reserach Article, etc...
- **aux_issn**: 175.690 nomes de veículos de publicação com seus respectivos ISSN, oriundos da base OpenAlex.
- **aux_vehicles**: 37.685 registros de veículos de publicação, tais como conferências, revistas científicas e outros.
- **aux_paises**: 974 formas de escrita de países e sua respectiva forma de escrita em português, por extenso e com sigla de 3 caracteres (ISO 3166-1 alpha-3).

Muitas produções científicas são oriundas de comunicações em simpósios, congressos e eventos e este conteúdo, quando oriundo do currículo Lattes, embora riquíssimo, traz associadas 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 (aux_vehicles). 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 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, em geral em planilhas. É o caso de dados de pessoal, fornecidos pela COGEP, área de Recursos Humanos da Fiocruz e de dados de áreas do conhecimento do CNPq em que cada publicação foi classificada pelas Unidades da Fiocruz.

Uma *Application Programming Interface* (API), é um conjunto de regras e protocolos que possibilita que diferentes softwares interajam entre si para troca de informações acessando mutuamente recursos como funções, dados ou serviços.

Scripts in-house são programas de computador desenvolvidos pela equipe do Observatório para processar os dados. Os scripts foam desenvolvidos utilizando a linguagem de programação Python versão 3.11 (https://www.python.org).

PostgreSQL é um sistema de gerenciamento de banco de dados relacional (SGBDR) de código aberto e amplamente utilizado para armazenar, organizar e gerenciar dados. O PostgreSQL é conhecido por sua confiabilidade, flexibilidade e suporte a padrões técnicos abertos, sendo capaz de lidar com altas cargas de trabalho (https://www.postgresql.org).

[2] https://www.thevantagepoint.com

JavaScript Object Notation (JSON) é um formato de arquivos baseado em texto para representar dados. É comumente utilizado como formato padrão para interoperabilidade em sistemas e aplicativos da Web (https://www.w3.org/TR/json-ld11).

IDENTIFICAÇÃO DE UNIDADES DA FIOCRUZ E SEUS PARCEIROS

Para construir os dicionários de afiliações (27 unidades e 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 então 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* Vantage Point®. Este processo se aplica às seis primeiras bases analisadas, aplicando os dicionários construídos a partir das instituições que cada autor declarou na publicação como afiliações.

Para a sétima base, o currículo Lattes, a afiliação considerada foi aquela assinalada pela 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 o Currículo de servidores, que por tanto, asseguram sua afiliação com a Fiocruz.

Pretende-se no futuro realizar a integração entre o Observatório e a Cogepe, o que permitirá em novas versões do dashboard, capturar o histórico de migração dos servidores entre as unidades.

Por fim, é importante mencionar que 12.780 produções científicas estão agrupadas com o rótulo de unidade "Fiocruz". Este número se refere a produções para as quais não identificamos, 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 vinculo, assim é possível uma melhor identificação das publicações de cada unidade da Fiocruz.

PADRONIZAÇÃO DOS AUTORES

Nesta etapa, após todas as tentativas descritas de normalização dos nomes de autores, (Vinculação ao ORCID, vinculação ao IdLattes, Vinculação ao Wos_ID e Scopus_ID, além de dicionários para normalização dos nomes por unidade), optou-se por considerar apenas na visualização final os autores servidores da Fiocruz, haja vista que estes são em sua ampla maioria os autores objeto de interesse das publicações institucionais.

Na coleta do currículo Lattes, não foram considerados os co-autores das publicações. Novas tentativas e outros testes para normalização utilizando algoritmos de reconhecimento como *strings* com busca aproximada através da distância de Levenshtein, estão em curso, mas na intenção de preservar a integridade do dado e garantir a não vinculação de publicações de homônimos, optou-se por este formato de visualização.

RECURSOS EXTERNOS

OPENALEX

A base OpenAlex é um catálogo aberto e global de pesquisa, criada pela organização sem fins lucrativos OurResearch, assim nomeada em homenagem à antiga biblioteca de Alexandria. Obtivemos a partir da OpenAlex uma lista de 175.691 veículos de publicação com seus respectivos ISSN e eISSN. Com estes dados foi construído um dicionário de ISSN.

ORCID

A base *Open Researcher and Contributor* ID (ORCID) é uma base global, mantida por uma organização sem fins lucrativos que fornece um identificador digital persistente capaz de distinguir pesquisadores que possuem um ORCID. Obtivemos, utilizando a API ORCID, 34.510 ORCIDs de pesquisadores que indicaram na base ORCID afiliação com alguma das unidades da Fiocruz. Estes dados foram cruzados com os dados de pessoal da COGEPE para indicar ORCIDs dos servidores ativos e inativos, quando disponível. Em um universo de 7.140 servidores, foi possível assinalar 1.453 ORCIDs. Os nomes foram buscados na API (Application Programming Interface) do ORCID[3].

LATTES ID

O currículo Lattes é uma base de currículos brasileira criada e mantida pelo CNPq. Cada pesquisador com um currículo Lattes cadastrado, é identificado por um número de 16 dígitos, o Lattes ID. O Lattes ID, semelhante ao ORCID, é também um identificador único que permite distinguir pesquisadores, em sua maioria brasileiros. Em um universo de 7.140 servidores ativos e inativos, foi possível assinalar 2.260 Lattes ID, o que possibilita uma normalização dos nomes e junção das publicações de mesma autoria.

CONVERSÃO DE FORMATO A PARTIR DAS BASES DE DADOS

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 DOAJ (https://doaj.org).

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

[3] https://info.orcid.org/documentation/features/public-api

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.

Importante salientar que, enquanto as demais bases são orientadas à publicação, o currículo Lattes é orientado 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: 2008 a 2023.

Ao converter os registros de cada base, alguns critérios de exclusão são aplicados:

Registros sem títulos
Títulos com menos de 10 caracteres
Títulos que contenham as palavras: *"author correction:", "author's reply", "comment on", "correction:", "corrigendum", "re:", "response to"*
Registros sem ano de publicação
Registros com ano de publicação diferente do intervalo 2008-2023
Registros sem autores

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.

LIMPEZA 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. As estratégias de comparação do dado de um registro com os dados dos dicionários são: busca exata, busca aproximada utilizando lógica fuzzy, busca aproximada por distância de Levenshtein. Com esta estratégia conseguimos “harmonizar” os dados estabelecendo uma convergência de nomes padronizados para a 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, 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 abrigas estes dados, tendo como sistema gerenciador de bancos de dados (SGBD), o PostgreSQL[4], 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).

A primeira informação inserida de dados, são os autores oriundos da COGEP com suas respectivas unidades de afiliação e número de ORCID, tanto para autores COGEP, quanto para os demais autores buscados via API ORCID. Esta inserção preliminar visa evitar duplicidades de autores a partir da checagem prévia no banco de dados pela sua existência quando inserindo uma publicação em que haja declarados ORCID como indicador persistente. 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 Curriculo Lattes.

Ao inserir um registro vindo de uma determinada base, o sistema checa se a publicação já existe no banco, para evitar sua inserção em duplicidade. 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

[4] https://www.postgresql.org

base de origem é atualizada, acrescendo a nova origem.

ENRIQUECIMENTO DOS DADOS COM ÁREAS DO CONHECIMENTO DO 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, haja vista que esta é consolidada, amplamente conhecida e utilizada por todos os pesquisadores.

Essa informação foi obtida de duas formas diferentes:

1) Diretamente das unidades ou da coleta do Currículo Lattes. Na Câmara Técnica de Pesquisa, foi acordado com todos os diretores das Unidades da Fiocruz que, cada Unidade faria a categorização de uma amostra (pelo menos os últimos 5 anos de produção) das suas publicações com a tipologia documental artigo. Essa categorização foi vinculada as áreas de conhecimento do CNPq, que é referência nacional na avaliação da produção científica. O sistema possui quatro níveis hierárquicos que permitem uma análise detalhada da área de conhecimento de cada publicação. Esses níveis, para este indicador, são entendidos como: Grande área, Área, Subárea e Especialidade, nível 1,2,3 e 4 respectivamente.

2) Diretamente da coleta das informações preenchidas pelo autor em seu Currículo Lattes.

Para ambas as formas de obtenção do dado, 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 e está, de acordo com o CNPq, é uma Área dentro da Grande área Ciências Humanas, por inferência estas correções foram feitas. 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,* pois será formado um grupo de trabalho para debate e inserção destas. Como exemplos pode-se citar: Bioinformática, Tecnologias Assistidas, Direito Sanitário, dentre outras sugestões.

VIEW CONSOLIDADA

Com os dados completamente inseridos e enriquecidos, estabeleceu-se um alto volume de dados e relações. Neste cenário, a busca por informação feita no banco de dados via *query,* tornou-se custosa em termos de desempenho computacional. Para mitigar este problema, foi criada uma *visão consolidada*, que é um objeto de banco de dados que guarda o resultado prévio de uma *query* em uma fonte de informação à parte que pode ser buscada de maneira mais rápida, aumentando o desempenho e reduzindo o tempo das consultas.

DASHBOARD

A partir da consulta à visão consolidada, os dados puderam ser indexados para tornarem-se indicadores. A indexação foi realizada utilizando o *software* Elasticsearch[5], o qual provê mecanismos para que outro *software* desta mesma suíte, o Kibana possa apresentar de forma gráfica, os indicadores selecionados. A apresentação em forma de *dashboard*, ou painel, é interativa e permite ao usuário filtrar, selecionar e até mesmo baixar os dados (de forma limitada).

OBSERVAÇÕES

Todos os *scripts* e dados utilizados no *pipeline* aqui descrito estão depositados em repositório institucional (GitLab).

O resultado deste trabalho, não obstante a informação disponibilizada ser fruto de um método razoavelmente bem estabelecido ao longo dos anos pelo Observatório, pode estar sujeito a eventuais incompletudes ou inexatidões. Entretanto, o modelo adotado, permite melhorias incrementais, muitas delas já planejadas e/ou em andamento

ANEXO I

Lista de referência compilada pelo Observatório C,T&I em Saúde da Fiocruz para utilização na coleta de dados nas bases.

("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 "COC-Fiocruz" 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

[5] https://www.elastic.co/pt

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 "FIOCRUZ/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")

ANEXO II

Padronização da tipologia documental.

key	value
research article	Artigo
review article	Artigo
Article	Artigo
article-commentary	Artigo
Article in press	Artigo
Artigo	Artigo
Artigo|nan	Artigo
nan|Artigo	Artigo
research-article	Artigo
Review	Artigo
review-article	Artigo
article	Artigo
article in press	Artigo
artigo	Artigo
artigo|nan	Artigo
nan|artigo	Artigo
research-article	Artigo
review	Artigo
review-article	Artigo
Review article	Artigo
Review-article	Artigo
artigo nan	Artigo
nan artigo	Artigo
Classical Article	Artigo
Comparative Study	Artigo
Historical Article	Artigo
Journal Article	Artigo
Journal Article (Default value when no more descriptive PT is provided or assigned)	Artigo
Review	Artigo
Scientific Integrity Review	Artigo
Systematic Review	Artigo
Twin Study	Artigo
Validation Study	Artigo
Adaptive Clinical Trial	Artigo
Clinical Study	Artigo
Clinical Trial	Artigo
Clinical Trial, Phase I	Artigo
Clinical Trial, Phase II	Artigo
Clinical Trial, Phase III	Artigo
Clinical Trial, Phase IV	Artigo
Clinical Trial Protocol	Artigo
Clinical Trial, Veterinary	Artigo
Controlled Clinical Trial	Artigo
Equivalence Trial	Artigo
Evaluation Study	Artigo
Observational Study	Artigo
Observational Study, Veterinary	Artigo
Pragmatic Clinical Trial	Artigo
Randomized Controlled Trial	Artigo
Randomized Controlled Trial, Veterinary	Artigo
Meta-Analysis	Artigo
case-report	Artigo
case-report	Artigo
case report	Artigo
relato de caso	Artigo
Case Reports	Artigo
Legal Case	Artigo

key	value
Multicenter Study	Artigo
Article;Article	Artigo
artigo comunicação em evento	Comunicação em evento
artigo comunicação em evento|nan	Comunicação em evento
Artigo|Comunicação em evento	Comunicação em evento
Artigo|Comunicação em evento|nan	Comunicação em evento
Comunicação em evento	Comunicação em evento
Comunicação em evento|Artigo	Comunicação em evento
Comunicação em evento|Artigo|nan	Comunicação em evento
Comunicação em evento|nan	Comunicação em evento
Conference paper	Comunicação em evento
Meeting Abstract	Comunicação em evento
nan|Comunicação em evento	Comunicação em evento
Proceedings Paper	Comunicação em evento
artigo|comunicação em evento	Comunicação em evento
artigo|comunicação em evento|nan	Comunicação em evento
comunicação em evento	Comunicação em evento
comunicação em evento|artigo	Comunicação em evento
comunicação em evento|artigo|nan	Comunicação em evento
comunicação em evento|nan	Comunicação em evento
conference paper	Comunicação em evento
meeting abstract	Comunicação em evento
nan|comunicação em evento	Comunicação em evento
proceedings paper	Comunicação em evento
nan comunicação em evento	Comunicação em evento
comunicação em evento	Comunicação em evento
Clinical Conference	Comunicação em evento
Congress	Comunicação em evento
Consensus Development Conference	Comunicação em evento
Consensus Development Conference, NIH	Comunicação em evento
Papers presented at events	Comunicação em evento
Data paper	Outros
data paper	Outros
Introductory Journal Article	Editorial
editorial	Editorial
Editorial	Editorial
Editorial Material	Editorial
editorial	Editorial
editorial	Editorial
editorial material	Editorial
Editorial	Editorial
Editorial	Editorial
Book	Livro ou capítulo de livro
Book chapter	Livro ou capítulo de livro
Book part	Livro ou capítulo de livro
book	Livro ou capítulo de livro
book chapter	Livro ou capítulo de livro
book part	Livro ou capítulo de livro
livro ou capítulo de livro	Livro ou capítulo de livro
Festschrift	Livro ou capítulo de livro
Nan	Não identificado
undefined	Não identificado
nan	Não identificado
undefined	Não identificado
article-commentary	Outros
article commentary	Outros
article commentary	Outros
Biographical-Item	Outros
biographical-item	Outros
biographical item	Outros
biografia	Outros
Autobiography	Outros

key	value
Biography	Outros
brief-report	Outros
press-release	Outros
rapid-communication	Outros
brief-report	Outros
press-release	Outros
rapid-communication	Outros
brief report	Outros
press release	Outros
rapid communication	Outros
book-review	Outros
Book Review	Outros
book-review	Outros
book review	Outros
book review	Outros
Note	Outros
note	Outros
nota	Outros
addendum	Outros
Outros	Outros
addendum	Outros
outros	Outros
Dataset	Outros
Government Publication	Outros
Guideline	Outros
Interactive Tutorial	Outros
Interview	Outros
Lecture	Outros
Legislation	Outros
News	Outros
Newspaper Article	Outros
Overall	Outros
Patient Education Handout	Outros
Periodical Index	Outros
Personal Narrative	Outros
Portrait	Outros
Practice Guideline	Outros
Research Support, American Recovery and Reinvestment Act	Outros
Research Support, N.I.H., Extramural	Outros
Research Support, N.I.H., Intramural	Outros
Research Support, Non-U.S. Gov't	Outros
Research Support, U.S. Gov't, Non-P.H.S.	Outros
Research Support, U.S. Gov't, P.H.S.	Outros
Technical Report	Outros
Video-Audio Media	Outros
Webcast	Outros
Bibliography	Outros
Collected Work	Outros
Dictionary	Outros
Directory	Outros
Duplicate Publication	Outros
Electronic Supplementary Materials	Outros
Expression of Concern	Outros
Letter	Outros
letter	Outros
Prefácio	Outros
prefacio	Outros
preface	Outros
abstract	Outros
abstract	Outros
resumo	Outros
English Abstract	Outros

key	value
Short survey	Outros
short survey	Outros
Technical Manuals and Procedures	Outros
Technical Manuals and Procedures;Reference term	Outros
Corrected and Republished Article	Remove
Preprint	Remove
Address	Remove
Comment	Remove
correction	Remove
correction	Remove
Erratum	Remove
erratum	Remove
Published Erratum	Remove
Retracted	Remove
retracted	Remove
Retracted Publication	Remove
Retraction of Publication	Remove