UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
BACHARELADO EM CIÊNCIAS BIOLÓGICAS

Laís Duarte Iesbich

DESENVOLVIMENTO E IMPLEMENTAÇÃO DE UMA FERRAMENTA DE AUTOMAÇÃO EM PYTHON PARA ANÁLISE DE CURRÍCULOS LATTES NO NÚCLEO DE BIOINFORMÁTICA DO HOSPITAL DE CLÍNICAS DE PORTO ALEGRE

Porto Alegre
2024

Laís Duarte Iesbich

DESENVOLVIMENTO E IMPLEMENTAÇÃO DE UMA FERRAMENTA DE AUTOMAÇÃO EM PYTHON PARA ANÁLISE DE CURRÍCULOS LATTES NO NÚCLEO DE BIOINFORMÁTICA DO HOSPITAL DE CLÍNICAS DE PORTO ALEGRE

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Ciências Biológicas com ênfase em bioinformática na Universidade Federal do Rio Grande do Sul.

Orientadora: Dra. Giovanna Câmara Giudicelli

Porto Alegre

2024

FICHA CATALOGRÁFICA

CIP - Catalogação na Publicação

Iesbich, Laís Duarte
DESENVOLVIMENTO E IMPLEMENTAÇÃO DE UMA FERRAMENTA DE AUTOMAÇÃO EM PYTHON PARA ANÁLISE DE CURRÍCULOS LATTES NO NÚCLEO DE BIOINFORMÁTICA DO HOSPITAL DE CLÍNICAS DE PORTO ALEGRE / Laís Duarte Iesbich. -- 2024.
24 f.
Orientadora: Giovanna Câmara Giudicelli.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Biociências, Bacharelado em Ciências Biológicas, Porto Alegre, BR-RS, 2024.

1. Bioinformática. 2. Automação. 3. Currículo Lattes. 4. Suporte à pesquisa. 5. Python. I. Giudicelli, Giovanna Câmara, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

FOLHA DE APROVAÇÃO

Laís Duarte Iesbich

DESENVOLVIMENTO E IMPLEMENTAÇÃO DE UMA FERRAMENTA DE AUTOMAÇÃO EM PYTHON PARA ANÁLISE DE CURRÍCULOS LATTES NO NÚCLEO DE BIOINFORMÁTICA DO HOSPITAL DE CLÍNICAS DE PORTO ALEGRE

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Ciências Biológicas com ênfase em bioinformática na Universidade Federal do Rio Grande do Sul.

Orientadora: Dra. Giovanna Câmara Giudicelli

Porto Alegre
Agosto de 2024

BANCA EXAMINADORA:

_____

Dra. Giovanna Câmara Giudicelli, pesquisadora no Núcleo de Bioinformática do Hospital de Clínicas de Porto Alegre.

_____

Prof. Dr. Michael Everton Andrades, pesquisador do Centro de Pesquisa Experimental do Hospital de Clínicas de Porto Alegre.

_____

Profa. Dra. Ursula da Silveira Matte, professora adjunta do Departamento de Genética da Universidade Federal do Rio Grande do Sul.

AGRADECIMENTOS

Agradeço aos meus pilares, minha família, ao meu pai e ao Barney, pelo apoio constante e amor incondicional. Sem essa base, minha trajetória não seria possível.

Às minhas amigas biólogas Andressa, Bianca, Marina, Mariana, Luiza e minha gêmea Duda, parceiras de perrengue nas saídas de campo, nas filas quilométricas do RU no Campus do Vale e grupo de estudos. Sou grata pela amizade e pelo apoio fundamental em todos os desafios e conquistas que vivemos juntas. Agradeço ao meu gêmeo Raul, que foi meu parceiro ao iniciar na bioinformática e sempre incentiva todos ao seu redor a crescerem como profissionais.

Às minhas amigas Gabriela e Isadora, meus pilares na vida pessoal há muitos anos e profissionais corporativas que admiro e me inspiro. Agradeço por estarem sempre ao meu lado.

E, finalmente, gostaria de expressar minha profunda gratidão à minha orientadora, Giovanna, que foi a melhor pessoa que eu poderia ter ao meu lado nesta etapa. Este projeto não teria sido o mesmo sem os métodos, orientações e as inúmeras reuniões que tivemos. A dedicação, o acolhimento e o suporte da Gio foram essenciais para o desenvolvimento e o sucesso deste trabalho.

RESUMO

O Currículo Lattes é crucial no meio acadêmico como ferramenta de avaliação, fornecendo uma visão padronizada e atualizada das atividades e produções dos pesquisadores. Utilizado por instituições e agências de fomento, o Lattes facilita a análise objetiva de publicações, orientações e outras contribuições, garantindo transparência e equidade na seleção e concessão de bolsas e promoções. O Hospital de Clínicas de Porto Alegre (HCPA), assim como outras instituições, utiliza o Lattes para validar a produção científica dos pesquisadores em diversos editais institucionais, como o Programa Institucional de Bolsas de Iniciação Científica (PIBIC). O desenvolvimento de ferramentas de suporte à pesquisa é essencial para acompanhar a inovação tecnológica e oferecer um suporte eficaz à tomada de decisão pela comissão de avaliação deste edital. Dada a necessidade de suporte tecnológico para a avaliação, este trabalho visou aprimorar o processo de análise dos currículos Lattes no HCPA, desenvolvido por meio do Núcleo de Bioinformática (NBioinfo). A aprimoração na avaliação foi alcançada por meio de um script de automação. Este código foi projetado para ler os currículos em formato XML, extraindo os arquivos diretamente da plataforma Lattes, e compilar de forma ágil as informações dos pesquisadores em um documento consolidado e fácil de manipular. Desenvolvido em linguagem Python e utilizando a biblioteca Pandas, o script é altamente eficiente para automação. O Python oferece uma base sólida para automação e manipulação de dados, enquanto a biblioteca Pandas facilita a leitura, transformação e análise de grandes volumes de informações, atendendo perfeitamente às necessidades do projeto. A aplicação prática deste código pelo NBioinfo foi implementada para o Edital nº 04/2024 PIBIC/HCPA. A ferramenta demonstrou eficiência para a comissão avaliadora, substituindo o processo manual de conferência de currículos.

Palavras-chave: Plataforma Lattes; Avaliação acadêmica; Suporte à pesquisa; Inovação tecnológica; Gerenciamento de Dados.

ABSTRACT

The Lattes Curriculum is crucial in the academic field as an evaluation tool, providing a standardized and updated view of researchers' activities and productions. Used by institutions and funding agencies, the Lattes facilitates objective analysis of publications, mentoring, and other contributions, ensuring transparency and fairness in the selection and awarding of scholarships and promotions. The *Hospital de Clínicas de Porto Alegre* (HCPA), like other institutions, utilizes the Lattes to validate researchers' scientific output in various institutional calls, such as the *Programa Institucional de Bolsas de Iniciação Científica* (PIBIC) . The development of research support tools is essential to keep up with technological innovation and provide effective support for decision-making by the evaluation committee. Given the need for technological support in evaluation, this work aimed to enhance the process of analyzing Lattes curriculum at HCPA, developed through the Bioinformatics Core (NBioinfo). This enhancement was achieved through an automation script. This code was designed to read Lattes curriculum in XML format, which can be directly extracted from the Lattes platform, and to compile researchers' information quickly into a consolidated and easily manageable document, such as an Excel spreadsheet. Developed in Python and using the Pandas library, the script is highly efficient for automation. Python provides a solid foundation for automation and data manipulation, while the Pandas library facilitates the reading, transformation, and analysis of large volumes of information, perfectly meeting the project's needs. The practical application of this code by NBioinfo was implemented for Edital No. 04/2024 PIBIC/HCPA. The tool demonstrated its efficiency to the evaluation committee, replacing the manual process of reviewing curricula.

Keywords: Lattes Platform; Academic Evaluation; Research Support; Technology innovation; Data management.

1. INTRODUÇÃO

A Plataforma Lattes (https://lattes.cnpq.br), desenvolvida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), integra bases de dados de currículos, grupos de pesquisa e instituições em um sistema único de informações. Este sistema tornou-se o padrão nacional para registrar as trajetórias acadêmica e profissional de estudantes e pesquisadores no Brasil, sendo amplamente adotado por entidades de apoio à pesquisa, universidades e institutos. A plataforma é essencial para a avaliação de mérito e competência em solicitações de financiamento na área de ciência e tecnologia, devido à sua extensa base de dados, confiabilidade e amplitude. Informações detalhadas sobre formação acadêmica, publicações, orientações e participação em projetos são exemplos de dados presentes nos currículos. Seleção de bolsas, financiamento de projetos e contratações acadêmicas realizam avaliações baseados em currículos Lattes de pesquisadores, podendo contar com outros processos para complementar a avaliação (CNPq, 2024).

Além da Plataforma Lattes, o Qualis tem papel fundamental na participação de processos de avaliação e classificação de produção acadêmica, pois fornece uma métrica de qualidade para as publicações científicas registradas nos currículos dos pesquisadores. O Qualis Periódicos é um sistema desenvolvido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) com o objetivo de classificar e avaliar a produção científica dos programas de pós-graduação no Brasil. Para que um periódico seja classificado pelo Qualis, é necessário que ele tenha recebido produções no ano-base e que essas produções sejam informadas pelos programas de pós-graduação no módulo Coleta da Plataforma Sucupira (CAPES, 2023). A classificação dos periódicos no Qualis é realizada pelos comitês de consultores de cada área de avaliação, seguindo critérios previamente definidos e aprovados pelo Conselho Técnico-Científico da Educação Superior (CTC-ES). Esses critérios estão disponíveis nos Documentos de Área da CAPES e são projetados para refletir a importância e a qualidade dos diferentes periódicos para cada área específica. Os periódicos são classificados em diferentes estratos, como A1 ao A4, B1 a B4, e C, sendo o A1 o mais alto padrão de qualidade (CAPES, 2023).

O Hospital de Clínicas de Porto Alegre (HCPA) é uma das instituições que faz uso da Plataforma Lattes para comprovar produção científica de seus pesquisadores para que eles possam, por exemplo, concorrer a bolsas de iniciação científica (IC). O Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do HCPA surgiu em 1994 e tem como objetivos promover o desenvolvimento científico e introduzir os estudantes à prática da pesquisa acadêmica. Para isso, o HCPA recebe cotas de bolsas de IC de agências de fomento como CNPq e FAPERGS, além de oferecer apoio financeiro institucional próprio. Essas bolsas são distribuídas entre alunos de graduação da área da saúde, que são recomendados por pesquisadores da instituição que possuem uma produção científica destacada (HCPA, 2024).

Os editais das bolsas PIBIC/HCPA possuem diferentes critérios de avaliação, que são julgados de forma parcialmente manual pela Comissão de Avaliação de Bolsas de Iniciação Científica, constituída por funcionários contratados do HCPA. A análise inicial do edital consiste em avaliar manualmente os dados registrados no currículo Lattes dos pesquisadores que se inscrevem no edital. Ao longo dos anos, identificou-se que analisar manualmente currículos Lattes de muitos pesquisadores não era a forma mais eficiente de se realizar essa tarefa. Devido às características do processo, como a repetitividade, a suscetibilidade a erros causados por falhas humanas, a estruturação padrão encontrada nos currículos, e a análise de dados, ele se torna um processo com viabilidade de automação (WEWERKA; REICHERT, 2020). A plataforma Lattes permite a extração de currículos em linguagem *Extensible Markup Language* (XML), a qual oferece diversas vantagens para automação de processos. A linguagem XML organiza os dados de maneira hierárquica através do uso de “tags”, facilitando a leitura e interpretação por diferentes sistemas (PACKER et al., 2014).

Ao pensar que a linguagem de programação pode estar distante do profissional biólogo, o XML oferece uma visão entre esses mundos aparentemente distintos. O XML utiliza uma estrutura semelhante à de uma árvore, onde o elemento "root" representa o ponto inicial e mais abrangente dos dados. Assim como os galhos se estendem e as folhas se interligam em uma árvore, o XML estrutura os dados em hierarquias distintas por meio de elementos (MUNZERT et al., 2014).

Cada nó da árvore representa informações precisas que são acessadas de maneira organizada e lógica.

Devido a problemática de uma tarefa manual e a capacidade de uma solução tecnológica, iniciou-se o desenvolvimento de uma ferramenta capaz de ler e extrair as informações necessárias a fim de avaliar os pesquisadores inscritos para o Edital nº 04/2024 PIBIC/HCPA. Existem ferramentas disponíveis online com funcionalidades semelhantes para a extração de informações de currículos Lattes, como o ScriptLattes (MENA-CHALCO; CESAR-JR, 2005) e o LucyLattes (TIEPPO, 2019). O ScriptLattes extrai informações do Lattes em formatos HTML, enquanto o LucyLattes lê essas informações a partir de arquivos XML, sendo este o script que mais se assemelha à biblioteca usada e às informações extraídas para este trabalho. Apesar de haver essas ferramentas disponíveis, elas ainda não atendem completamente às necessidades de avaliação exigidas pela diretoria de pesquisa.

Em 2022, um script com o mesmo propósito foi desenvolvido propriamente para o HCPA. Embora eficiente na época, o script tornou-se obsoleto e desatualizado. Consequentemente, foi necessário desenvolver uma nova solução que atendesse às necessidades atuais do HCPA, garantindo compatibilidade e funcionalidade. Para atender a essa necessidade, a comissão do Edital nº 04/2024 PIBIC/HCPA recorreu ao Núcleo de Bioinformática (NBioinfo) em busca do apoio tecnológico necessário para criar uma ferramenta para o edital atual, mas que também pudesse ser utilizada de forma contínua em outros editais.

Integrado à Unidade de Bioestatística e Análise de Dados, o NBioinfo desempenha um papel fundamental na pesquisa, colaboração e suporte nas áreas de bioinformática e biologia computacional. O NBioinfo visa gerar e compartilhar conhecimento científico de excelência, tanto básico quanto aplicado, capacitar profissionais qualificados e apoiar pesquisadores na criação de soluções computacionais para análise de dados.

Portanto, a automação proposta no presente trabalho não apenas otimizaria o processo de avaliação dos pesquisadores para o Edital nº 04/2024 PIBIC/HCPA, mas também representa um avanço significativo na eficiência e precisão das

análises, alinhando-se às demandas atuais por processos mais rápidos e menos suscetíveis a erros humanos. Isso é fundamental para garantir a qualidade das iniciativas de pesquisa no contexto do HCPA.

2. OBJETIVOS

2.1. Objetivo geral

Desenvolver e implementar script em linguagem Python capaz de extrair dados de currículos Lattes, automatizando a avaliação e proporcionando melhoria no processo.

2.2. Objetivos específicos

- Atender as especificações do Edital nº 04/2024 PIBIC/HCPA;
- Manipular e organizar os dados extraídos de forma eficiente, garantindo precisão de informações para a comissão avaliadora;
- Reduzir o tempo manual gasto na análise dos currículos;
- Garantir que o script seja compatível com diferentes sistemas operacionais para ampliar o seu uso e acessibilidade.

3. METODOLOGIA

3.1. Necessidades do Edital nº 04/2024 PIBIC/HCPA

Diversas especificações foram feitas pela comissão avaliadora do edital para que o código atendesse às necessidades do edital:

- **Analisar um Período Específico**: O período de avaliação do edital foi de 2019 a 2023. No entanto, o código deveria ser facilmente modificável para analisar outros períodos específicos conforme necessário.

- **Versatilidade para Análises Estendidas**: Considerando a inclusão de candidatas que gestaram durante o período de avaliação, o código deveria permitir a extensão do período de análise, considerando um ano extra para cada gestação.

- **Contabilizar artigos publicados:** Quantos artigos foram publicados pelos pesquisadores durante o período analisado, contabilizando somente aqueles artigos que apresentam DOI.

- **Quantos artigos publicados possuem classificação Qualis A:** Artigos com classificações A foram contabilizados de acordo com a base de classificações do quadriênio 2017-2020.

- **Orientações concluídas e em andamento:** O código precisava contabilizar separadamente as orientações concluídas e em andamento para mestrado, doutorado e IC.

3.2. Ferramentas utilizadas

Como uma boa dupla que funciona em diversos cenários, a linguagem para a automação deste projeto foi o Python (**Figura 1**), pois possui boas ferramentas e bibliotecas para trabalhar com XML. Segundo Brown (citado em Howe et al., 2015), Python *"é talvez mais complicada, mas também mais capaz: é adequada para tudo,*

desde automatizar pequenos conjuntos de instruções, até construir websites e aplicações completas".

Figura 1: Logotipo Python

Para o funcionamento de todo o código foram utilizados módulos e a biblioteca Pandas, cada um dos módulos é um programa Python que contém um grupo relacionado de funções que podem ser incorporadas em seus programas (SWEIGART, 2015), são eles, o xml.etree.ElementTree e os.py.

O Pandas (**Figura 2**) é uma biblioteca poderosa para trabalhar com XML, pois oferece uma estrutura de dados eficiente, conhecida como DataFrame, que facilita a manipulação, análise e limpeza de dados. Isso é especialmente útil ao lidar com arquivos de extensão XML, que podem ter estruturas complexas e aninhadas.

Figura 2: Logotipo Pandas

O Visual Studio Code (VS Code, **Figura 3**) é um software que fornece ferramentas abrangentes para escrever, testar e depurar código de maneira eficiente e organizada. Foi por meio desta ferramenta que o script de automação foi desenvolvido. Além disso, o software é compatível com diversos sistemas operacionais, tornando o script acessível a qualquer pessoa que precise executar a atividade. O presente trabalho foi realizado no sistema operacional Windows, mas pode ser rodado tanto em macOS quanto Linux.

Figura 3: Logotipo VS Code

4. RESULTADOS

4.1. Construção do código

Conforme a documentação da Python Software Foundation (2024), usada como referência para todo o código, a construção teve início com a chamada dos módulos e da biblioteca a ser usada. O código começa com a leitura de um currículo em formato XML usando as seguintes linhas:

```
import xml.etree.ElementTree as ET
import os
import pandas as pd

tree = ET.parse('curriculo.xml')
root = tree.getroot()
```

Essas linhas inicializam o processo de análise do arquivo XML, onde `ET.parse('curriculo.xml')` lê o currículo especificado, e `getroot()` obtém o elemento raiz do documento XML. Essa configuração permite que o script processe um currículo por vez, inserindo o caminho do arquivo XML.

A função `get` foi utilizada várias vezes ao longo do código para obter informações específicas do XML. Essa função facilita a navegação pelo documento e a extração de dados essenciais, como o nome do pesquisador e o número identificador do currículo. Essas informações foram formatadas em uma string para facilitar a leitura e está sendo armazenada na variável `Infos_Lattes`. Outra informação essencial para o edital, como o ano de doutorado, foi obtida de forma similar, seguindo a mesma lógica:

```
dados_curriculo = root
nome_completo_lattes = dados_curriculo.find('.//DADOS-GERAIS').get('NOME-COMPLETO')
numero_ID = dados_curriculo.get('NUMERO-IDENTIFICADOR')

Infos_Lattes = f"{nome_completo_lattes} - ID: {numero_ID}"
```

No requisito de produções do Currículo Lattes, a contagem de artigos publicados em periódicos é um dos indicadores importantes para a avaliação. A função `len ()` desempenha um papel fundamental nesse processo, oferecendo uma forma rápida e eficiente de contar os artigos, especialmente porque essa seção é uma das maiores em termos de conteúdo nos currículos Lattes.

Após contar a quantidade de artigos publicados e armazenar o valor na variável `total_artigos`, o código deve ler informações adicionais sobre cada artigo. Isso inclui acessar os dados básicos do artigo, onde são buscadas informações como o ano de publicação e a presença de DOI, obedecendo às necessidades do edital:

```
artigos_publicados = root.findall('.//ARTIGO-PUBLICADO')

total_artigos = len(artigos_publicados)

soma_artigos_intervalo = 0

for artigo in artigos_publicados:

    dados_basicos = artigo.find('DADOS-BASICOS-DO-ARTIGO')
    if dados_basicos is not None:
        doi = dados_basicos.get('DOI') or ""
        ano = int(dados_basicos.get('ANO-DO-ARTIGO'))

        if doi:
            total_artigos += 1
```

No processo de inscrição no edital, os pesquisadores foram instruídos a declarar se haviam tirado licença maternidade durante o período de avaliação de 2019 a 2023. Esta informação era essencial para possibilitar a adição de um ano extra de avaliação, garantindo equidade no processo de seleção. Em resposta a essa necessidade, o código deve ser suficientemente flexível para permitir alterações no período de análise. Além disso, essa flexibilidade assegura que o código possa ser adaptado para atender às exigências de futuros editais:

```
            if 2019 <= ano <= 2023:
                soma_artigos_intervalo += 1
```

Além da contagem dos artigos, é necessário verificar quantos desses artigos publicados possuem classificação Qualis. O código acessa a base Qualis para pesquisar pelo ISSN ou pelo nome da revista da publicação, a fim de atribuir a classificação ao artigo. A base Qualis é obtida diretamente no site do Sucupira, deve estar no formato xlsx e precisa estar previamente no mesmo diretório para que o código possa realizar a conferência com o currículo e, assim, contabilizar quantos artigos possuem classificações A:

```
def contar_artigos_por_estrato(xml, ano_inicio, ano_fim):

    Q = pd.read_excel('Qualis.xlsx')
    Q_dict = Q.set_index('ISSN')['Estrato'].to_dict()

    # artigos por estrato
    contagem_estratos = {'A1': 0, 'A2': 0, 'A3': 0, 'A4': 0}
```

Ao acessar a seção de orientações do Lattes, é possível observar como o XML organiza os dados em uma estrutura hierárquica com ramificações, similar a árvores. Para analisar todos todos os tipos de orientações, tanto concluídas quanto em andamento, foi necessário criar blocos dentro da função `contar_orientacoes`, detalhando cada tipo de orientação individualmente:

```
def contar_orientacoes(xml, ano_inicio, ano_fim):
    tree = ET.parse(xml)
    root = tree.getroot()

    orientacoes_por_tipo = {'Mestrado': 0 , 'IC': 0, 'TCC': 0, 'Doutorado': 0, 'Pos-Doutorado': 0}
    orientacoes_por_tipo_andamento = {'Mestrado': 0, 'IC': 0, 'TCC': 0, 'Doutorado': 0, 'Pos-Doutorado': 0}

    for orientacoes_concluidas in root.findall('.//ORIENTACOES-CONCLUIDAS'):
        for orientacao_concluida in orientacoes_concluidas.findall('.//ORIENTACOES-CONCLUIDAS-PARA-MESTRADO'):
            dados_orientacao = orientacao_concluida.find('DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO')
            if dados_orientacao is not None:
                ano = int(dados_orientacao.get('ANO'))
                if ano_inicio <= ano <= ano_fim:
                    orientacoes_por_tipo['Mestrado'] += 1
```

As orientações de Trabalho de Conclusão de Curso e Pós-Doutorado não foram contabilizadas no edital de referência. No entanto, o código já está preparado para incluir essas orientações em futuras avaliações. Foi necessário apenas ocultar os resultados referentes a esses campos nos resultados atuais, sem causar qualquer impacto negativo no processo.

Ao final do código foi criado um DataFrame, que são basicamente dados em tabelas com colunas específicas para cada resultado. A biblioteca Pandas fornece uma amplos métodos e funções para manipular dados em uma tabela. O DataFrame é preenchido usando a notação `df.loc[row, column]`, onde `row` é o índice da linha e `column` é o nome da coluna:

```
df = pd.DataFrame(columns=['Nome', 'ID', 'Avaliação entre', 'Ano Doutorado',
'Artigos Publicados', 'Artigos A1 a A4 do periodo analisado', 'Ori concluidas Mestrado',
'Ori concluidas Doutorado', 'Ori concluidas Pos-Doutorado', 'Ori concluidas IC', 'Ori concluidas TCC',
'Ori em andamento Mestrado', 'Ori em andamento Doutorado', 'Ori em andamento Pos-Doutorado',
'Ori em andamento IC', 'Ori em andamento TCC'])

df.loc[0, 'Nome'] = nome_completo_lattes
df.loc[0, 'ID'] = numero_ID
...
```

O DataFrame suporta exportação para diversos formatos, como Excel, Comma-Separated Values (CSV) e Structured Query Language (SQL). Devido a sua acessibilidade e familiaridade por muitos usuários, o Excel foi escolhido para a consolidação dos resultados. O fato de que muitos profissionais e acadêmicos já estão acostumados a trabalhar com Excel facilita a interpretação e a análise dos dados apresentados, além de garantir que os resultados possam ser compartilhados e compreendidos com facilidade:

```
excel_file = f"{nome_completo_lattes}.xlsx"
df.to_excel(excel_file, index=False)
```

4.2. Análise dos currículos Lattes

Os resultados de acordo com o tempo da execução do script de análise dos currículos Lattes dos participantes:

- O currículo mais extenso, considerando o mais complexo e volumoso, foi processado em **7 segundos**.

- O currículo menos extenso foi processado em **6.9 segundos**.

Esses resultados indicam que o tempo de análise pelo script é consistente, independentemente do tamanho do currículo Lattes avaliado, mantendo cerca de sete segundos por análise.

Durante o processo de análise dos currículos Lattes dos participantes, o código analisou cada currículo separadamente, gerando um arquivo Excel para cada pesquisador. Foi desenvolvido um arquivo final no formato Excel para compilar e visualizar os resultados obtidos. A **Tabela 1** ilustra a estrutura desse arquivo, que foi configurado para representar a organização dos dados coletados. Esta tabela foi criada como uma simulação, garantindo a confidencialidade dos nomes dos pesquisadores e dos resultados reais.

A tabela proporcionou uma visão estruturada e acessível dos dados processados, facilitando a interpretação e análise dos resultados gerados pelo código. A capacidade de ordenar, filtrar e aplicar pesos aos diferentes atributos dos pesquisadores, forneceu uma base sólida para análises estatísticas mais detalhadas. Isso permitiu que os membros da comissão do edital realizassem tomadas de decisões mais embasadas.

Tabela 1: Resultados compilados das Análises de Currículos Lattes por Rodadas de Execução do Código.

Nome	ID	Avaliação entre	Ano Doutorado	Artigos Publicados	Artigos A1 a A4 do período analisado	Ori concluídas Mestrado	Ori concluídas Doutorado	Ori concluídas IC	Ori em andamento Mestrado	Ori em andamento Doutorado	Ori em andamento IC
Pesquisador 1	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
Pesquisador 2	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
Pesquisador 3	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
Pesquisador 4	xx	2018 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
Pesquisador 5	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
Pesquisador 6	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx
…		…	…	…	…	…	…	…	…	…	…
Pesquisador 146	xx	2019 - 2023	xxxx	xx	xx	xx	xx	xx	xx	xx	xx

Fonte: Elaborada pela autora. Reprodução do arquivo.

5. CONCLUSÃO

Análises que antes dependiam de processos manuais, sujeitos a limitações de tempo e recursos humanos, foram significativamente otimizadas através da automação proporcionada pelo código desenvolvido para o Edital nº 04/2024 PIBIC/HCPA. Antes, um processo que era feito manualmente, podendo levar horas dependendo do currículo, foi simplificado com a nova ferramenta, reduzindo o tempo de processamento para apenas sete segundos por currículo. Essa mudança aumentou a velocidade do procedimento sem comprometer a precisão e a consistência das análises realizadas.

O script em Python desenvolvido mostrou-se capaz de ler e processar rapidamente currículos Lattes em formato XML, compilando informações necessárias. A implementação deste código no NBioinfo/HCPA permitiu a avaliação rápida e precisa de pesquisadores para a concessão de bolsas. O código não se restringe apenas ao edital referência, sendo flexível e adaptável para aplicação em futuros editais, avaliações de produtividade acadêmica e outras seleções realizadas pelo HCPA. Além disso, o script pode ser executado em diferentes sistemas operacionais, afirmando a flexibilidade do trabalho em diferentes ambientes, promovendo acessibilidade e usabilidade em diversas plataformas.

A criação deste script também destaca a importância da integração de tecnologias avançadas, como Python, na área da bioinformática e da gestão de processos acadêmicos. Este trabalho contribui no incentivo para o desenvolvimento de novas ferramentas automatizadas que possam abordar outras necessidades institucionais, incentivando uma cultura de inovação contínua no HCPA, demonstrando como a tecnologia pode ser uma aliada poderosa na gestão de informações e na promoção da eficiência institucional.

6. CONSIDERAÇÕES FINAIS

É crucial destacar que o script requer um ciclo contínuo de manutenção, atualização e aprimoramento após sua implementação inicial. Esse processo é fundamental para assegurar o funcionamento correto do código e para atender continuamente às necessidades do HCPA.

Documentar o código é essencial para que o trabalho desenvolvido seja facilmente compreendido, mantido e expandido por outras pessoas no futuro. No contexto deste TCC, além da apresentação detalhada na dissertação, foram preparados e entregues ao Núcleo de Bioinformática e à Unidade de Bioestatística e Análise de Dados do HCPA documentação com descrição minuciosa das etapas do processo, orientações para a execução e manutenção do código, e instruções para realizar atualizações. Esses materiais foram projetados para fornecer uma visão clara e completa do fluxo de trabalho e da funcionalidade do código. Mesmo após a conclusão deste projeto, qualquer novo usuário ou desenvolvedor poderá facilmente entender, utilizar e adaptar a ferramenta conforme necessário.

Como desenvolvedora, analisei duas possíveis melhorias para o código. A primeira melhoria seria permitir que o código processasse todos os arquivos XML localizados em um mesmo diretório de uma só vez, gerando uma única planilha ao longo do processo. Essa alteração aumentaria ainda mais a automação e eficiência do processamento dos dados. A segunda melhoria seria o desenvolvimento de uma interface mais amigável para usuários que não estão familiarizados com linhas de código. Uma interface gráfica permitiria aos usuários selecionar o arquivo XML do pesquisador e definir o período de análise de forma intuitiva e acessível. Essas melhorias visam tornar o processo mais eficiente e acessível, beneficiando uma gama mais ampla de usuários.

7. REFERÊNCIAS BIBLIOGRÁFICAS

CAPES. Qualis Periódicos. Plataforma Sucupira. Disponível em: https://sucupira.capes.gov.br/sucupira/public/index.jsf. Acesso em: 15 jul. 2024.

CITING AND LOGO. Pandas. Disponível em: https://pandas.pydata.org/about/citing.html. Acesso em: 1 jul. 2024.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. Sobre a plataforma Lattes. In: LATTES. Brasília, DF: CNPq, 2024. Disponível em: http://lattes.cnpq.br. Acesso em: 06 jul. 2024

HCPA. Área do Pesquisador. Bolsas PIBIC. Disponível em: https://sites.google.com/hcpa.edu.br/area-do-pesquisador/bolsas?authuser=0. Acesso em: 26 jul. 2024.

HOWE,A., et al. Pick up Python. *Nature*, 518, 125, 2015.

ICONS AND NAMES USAGE GUIDELINES. Pandas. Disponível em: https://pandas.pydata.org/about/citing.html. Acesso em: 1 jul. 2024.

MENA-CHALCO, J. P.; CESAR-JR, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, v. 15, n. 4, p. 31-39, 2009.

MUNZERT, S.; RUBBA, C.; MEIßNER, P.; NYHUIS, D. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons, Ltd, 2014.

PACKER, A., et al. Why XML? [online]. SciELO in Perspective. Disponível em: https://blog.scielo.org/en/2014/04/04/why-xml/. Acesso em: 03 Jun. 2024.

PYTHON SOFTWARE FOUNDATION. *Python Documentation*. Disponível em: https://docs.python.org/3/. Acesso em: 2 jul. 2024.

SWEIGART, A. *Automate the Boring Stuff with Python: Practical Programming for Total Beginners*. No Starch Press, 2015.

THE PYTHON LOGO. Python Software Foundation. Disponível em: https://www.python.org/community/logos/. Acesso em: 1 jul. 2024.

TIEPPO, R. LucyLattes: Extração automatizada de informações de currículos Lattes em XML. 2019. Disponível em: https://rafatieppo.github.io/post/2019_03_13_lucylattes/. Acesso em: 14 ago. 2024.

WEWERKA, J.; REICHERT, M. Towards quantifying the effects of robotic process automation. In: Proceedings of the IEEE International Enterprise Distributed Object Computing Workshops, 2020, online, p. 15. DOI: 10.1109/EDOCW49879.2020.00015.