Extrator de produções acadêmicas de pesquisadores do IFCE Campus Tianguá pela Plataforma Lattes.

Adonias Caetano de Oliveira[1], **Fabiano Tavares da Silva** [2]

[1] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Av. Tabelião Luiz Nogueira de Lima S/N – Tianguá – CE – Brasil

[2]Centro Universitário Farias Brito
R. Castro Monte, 1364 - Varjota, Fortaleza - CE - Brasil

{adonias.ifce, fabiano.bie}@gmail.com

Abstract. *The research coordinators from the IFCE constantly need to analyze the lattes curricula of their employees who work in the research. This manual analysis procedure for medium and large groups becomes a time-consuming and highly error-prone task. For this reason, this article reports the development of an extractor that automates the acquisition of curricula and the quantification of academic productions of the researchers from the campus Tianguá. The results show that it is an effective alternative in the acquisition of curricula, processing and calculation of the scientific production rate without the need for the user to manually download the curricula or solve the captcha.*

Resumo. *Os coordenadores de pesquisa do IFCE precisam constantemente analisar os currículos lattes de seus servidores que atuam na pesquisa. Esse procedimento manual de análise para médios e grandes grupos torna-se uma tarefa demorada e altamente suscetível a erros. Por esse motivo, este artigo relata o desenvolvimento de um extrator que automatiza a aquisição dos currículos e a quantificação das produções acadêmicas de pesquisadores do campus Tianguá. Os resultados mostram que ele é uma alternativa eficaz na aquisição dos currículos, processamento e cálculo da taxa de produção científica sem a necessidade do usuário baixar manualmente os currículos ou resolver o captcha.*

1. Introdução

No Brasil, a Plataforma Lattes[1] é uma das importantes fontes de acesso às informações sobre a vida pregressa e atual dos estudantes e pesquisadores do país. Essas informações são utilizadas pelas instituições acadêmicas e/ou grupos de pesquisa para elaborar relatórios, orientações e projetos de pesquisa. No intuito de avaliar, analisar ou documentar essas informações do grupo, ou mesmo de um campus de modo geral, os relatórios supracitados muitas vezes são produzidos de forma manual, considerando o currículo de cada membro [Mena-Chalco and Júnior 2013].

Este é o caso da Coordenadoria de Extensão e Pesquisa (CPE) do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) Campus Tianguá. Trimestralmente,

[1]Disponível em http://lattes.cnpq.br

esse setor precisa fazer coletas de dados dos pesquisadores do campus que estão cadastrados na plataforma NL. Esse ambiente de gerenciamento e sistematização de atividades de pesquisa foi desenvolvido pela Pró-Reitoria de Pesquisa, Inovação e Pós-Graduação (PRPI) com a meta de otimizar os fluxos de trabalho realizados pela comunidade de pesquisadores dos diversos campi do IFCE [PRPI 2021][2].

Entretanto, com a desatualização da ferramenta Extrator Lattes/PRPI desde junho de 2020, sem perspectiva de atualização, a pessoa responsável pela CPE de Tianguá precisa fazer uma análise manual de cada pesquisador do campus para calcular a taxa de produção científica. Mesmo com a informação semiestruturada, o procedimento de análise desses currículos para médios e grandes grupos torna-se uma tarefa demorada e altamente suscetível a erros [Mena-Chalco and Júnior 2013].

Diante disso, torna-se relevante a contribuição de pesquisas técnicas e científicas, como esta, sobre as potencialidades do desenvolvimento e aplicação de um protótipo baseado em Python no processo automático de extração de dados de currículos lattes de pesquisadores. Ademais, este artigo analisa e compara os resultados alcançados neste trabalho com sistemas equivalentes apontando benefícios, limitações e melhorias a serem realizadas numa versão final.

2. Fundamentação Teórica

A Plataforma Lattes, desenvolvida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), oferece acesso e padronização de informações acadêmicas e profissionais dos estudantes e pesquisadores atuantes em território brasileiro [Corrêa et al. 2017], além de permitir a integração de bases de currículos acadêmicos de pesquisadores, grupos de pesquisa e de instituições públicas e privadas em um único sistema de informações [BRITO et al. 2016, Branco et al. 2018]. Ela é utilizada por diversas instituições, como o próprio CNPq, Ministério da Ciência e Tecnologia, universidades e grupos de pesquisa [Corrêa et al. 2017].

Os chamados Currículos Lattes compõem uma base de dados caracterizada pela livre inserção de dados pelos próprios pesquisadores, ou seja, há currículos que são frequentemente atualizados, enquanto outros estão desatualizados há alguns anos. Geralmente, pesquisadores envolvidos com pesquisa acadêmica/aplicada em instituições de nível superior ou mesmo ligados a programas de Pós-Graduação tendem a manter seus currículos atualizados em função das avaliações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) ou para pleitear financiamentos de pesquisas e/ou bolsas junto às agências de financiamento [BRITO et al. 2016].

Adicionalmente, informações sobre os grupos de pesquisa científica e tecnológica em atividade no Brasil são mantidas pelo Diretório dos Grupos de Pesquisa (DGP[3]). São informações sobre recursos humanos (pesquisadores, estudantes e técnicos) que formam os grupos, especialidades do conhecimento, setores de aplicação envolvidos, linhas de pesquisa em andamento, capacidade científica, tecnológica, artística e as parcerias estabelecidas entre os grupos e as instituições, principalmente com as empresas do setor produtivo. Dessa maneira, torna-se uma ferramenta muito útil para descrever os limites e o perfil geral científico-tecnológico do país [Plataforma Lattes 2021].

[2]Disponível em http://prpi.ifce.edu.br/nl/app_Login/

[3]Disponível em http://lattes.cnpq.br/web/dgp

3. Trabalhos e softwares relacionados

Por meio do Google Scholar[4], foram selecionados três sistemas de coleta e extração de dados de currículos lattes no formato *eXtensible Markup Language* (XML).

3.1. Sistema da UNESP [Vidotti 2015]

Por meio de listas com os endereços permanentes dos currículos dos docentes de cada unidade da Universidade Estadual Paulista (UNESP), o sistema extrai o ID Lattes e baixa os currículos no formato XML. Contudo, é necessário que o operador resolva captcha para que o *download* seja autorizado pela Plataforma Lattes. Em setembro de 2015, o sistema foi capaz de calcular a quantidade de artigos produzidos e o aumento da quantidade de artigos (%) de 1091 docentes de seis unidades universitárias.

3.2. lucyLattes [Tieppo 2019]

O lucyLattes é um script em python, que faz a extração, a compilação, a organização dos dados dos currículos lattes em arquivos de texto, e a geração de um relatório simplificado, que proporcionam agilidade para a geração de informação. No entanto, ainda é necessário acessar e baixar manualmente os currículos desejados em formato XML.

3.3. UFOP Ativa [Lana 2019]

Esse sistema agrupa e organiza em um banco de dados as informações, provenientes da Plataforma Lattes, que estejam relacionadas aos servidores atuantes na Universidade Federal de Ouro Preto (UFOP). A partir de um arquivo *Comma-Separated Values* (CSV), que contém o nome, CPF, escolaridade, categoria, departamento, unidade e programa de pós-graduação de 1.653 servidores ativos em 2019, são obtidos os IDs Lattes. Os parâmetros CPF, o nome completo e a data de nascimento de uma pessoa (não sendo obrigatório o uso de todos) são usados nesse processo. Depois disso, ele consegue automaticamente baixar e extrair dados do currículo no formato XML através de um WebService SOAP, em um computador local da universidade. Esse acesso só foi possível devido a uma solicitação de acesso escrito em ofício devidamente assinado pelo dirigente máximo da UFOP e encaminhado à Presidência do CNPq.

4. Materiais e Métodos

Os procedimentos metodológicos adotados neste trabalho estão relacionados com aplicação de metodologias de Engenharia de Software (ES). Esta pesquisa foi organizada em cinco etapas, sendo elas: Revisão de Literatura, Requisitos do protótipo, Desenvolvimento do protótipo, Obtenção de resultados e Análise comparativa.

4.1. Requisitos do protótipo

Os requisitos do protótipo foram especificados com base em entrevistas com a coordenadora da CPE do IFCE Campus Tianguá e nos trabalhos relacionados.

A coordenadora precisa trimestralmente atualizar o indicador TAXA DE PRODUÇÃO CIENTÍFICA do campus com base nas informações sobre produções acadêmicas dos seus pesquisadores cadastrados na plataforma NL, as mesmas pessoas que estão cadastradas nos grupos de pesquisa.

[4]Disponível em https://scholar.google.com.br/?hl=pt

O cálculo dessa taxa utiliza a quantidade de pesquisadores, artigos publicados/apresentados, livros, capítulos de livros, notas técnicas, ações artísticas e culturais, registros ou depósitos de propriedades intelectual e outras produções afins. Em resumo, são calculadas por campus as quantidades de pesquisadores, de informações bibliográficas, técnicas e artísticas/culturais juntamente com a taxa científica, conforme a Equação 1.

$$\textbf{Taxa de Produção} = \frac{\text{Produção Científica}}{\text{Total de Pesquisadores}} \tag{1}$$

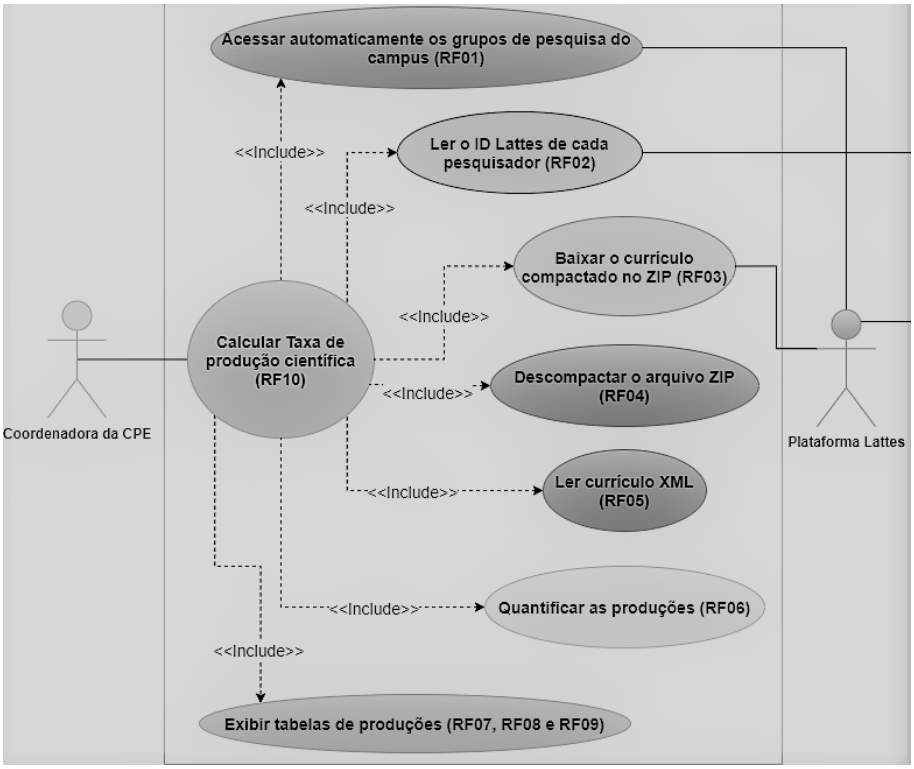

Figura 1. Diagrama de Casos de Uso

Usando a Linguagem de Modelagem Unificada (*Unified Modeling Language*, UML) foram especificados os casos de usos (requisitos) do protótipo, ver Figura 1. Foi usado o *software* Draw.io[5], que permite criar diversos diagramas da UML.

4.2. Desenvolvimento do protótipo

Neste trabalho, o extrator de dados do lattes está disponível apenas no ambiente Google Colaboratory porque facilita de forma rápida a implementação, os testes preliminares e uma validação dos resultados pela coordenadora. Por isso, ele é considerado apenas um protótipo neste artigo, já que é apenas um modelo preliminar para prova de conceito ou um Produto Viável Mínimo (MVP) se for executado diretamente pelo Colab.

Conforme ilustra a Figura 2, o projeto do protótipo foi organizado numa arquitetura lógica composta pelos seguintes módulos:

[5]Disponível em https://www.draw.io/

1. **Aquisição dos currículos:** Esse módulo é responsável por acessar as páginas dos grupos de pesquisas do campus usando *selenium-requests*[6], extrair os dados pessoais de cada pesquisador através do seu espelho RH, extrair o ID Lattes dos dados pessoais e acionar o *download* automático do currículo lattes compactado em arquivo ZIP usando o serviço 2Captcha[7].
2. **Pré-processamento dos currículos:** Esse módulo descompacta todos os arquivos ZIP baixados usando a biblioteca **zipfile**[8], ler os dados contidos em cada arquivo XML e armazena-os na estrutura dictionary do Python.
3. **Cálculo das produções:** Nesse módulo, cada tipo de atuação acadêmica por pesquisador é quantificada e armazenada em estruturas *data frames* da biblioteca pandas. Ao final, a taxa dessas atuações é determinada.
4. **Apresentação dos resultados:** Exibe a capacidade acadêmica do campus por meio de gráficos de barra.

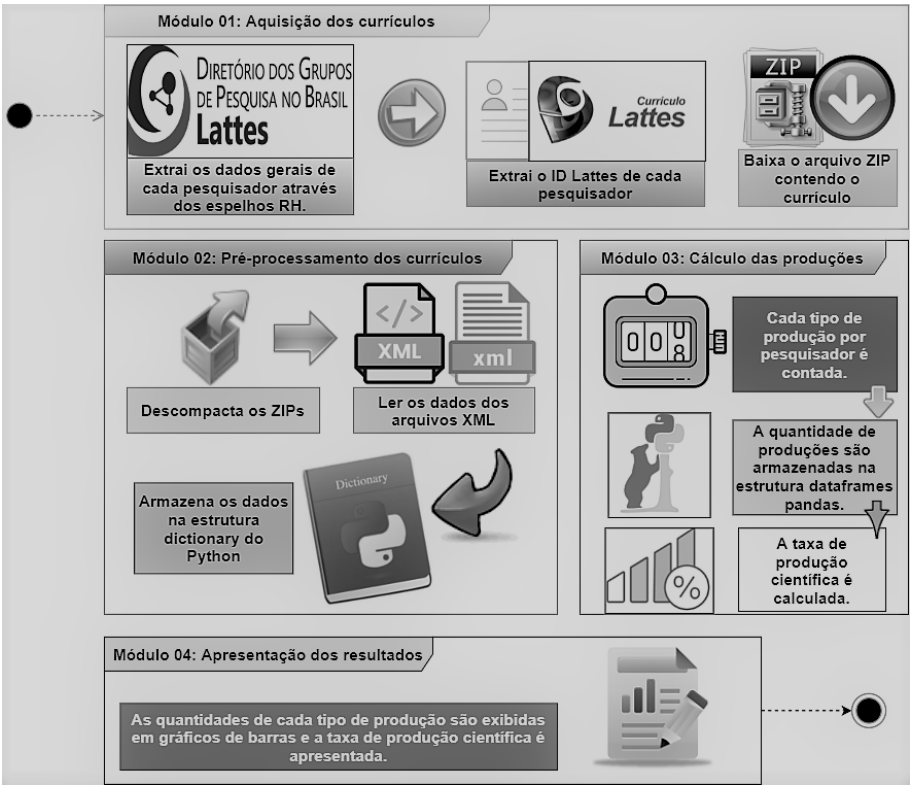

Figura 2. Arquitetura Lógica

A Plataforma Lattes permite imprimir o currículo em dois formatos, a saber, em PDF ou XML. Com base nas experiências relatadas pelos trabalhos relacionados, fazer a extração de informação dos arquivos XML é o método mais prático de identificar e quantificar as produções desejadas.

5. Resultados preliminares

Atualmente no campus, há 37 docentes ativos na pesquisa acadêmica e tecnológica, seja através de bolsas de iniciação científica, projetos de pesquisa voluntária, projetos de ex-

[6]Documentação disponível em https://pypi.org/project/selenium-requests/
[7]Disponível em https://2captcha.com
[8]Documentação disponível em https://docs.python.org/pt-br/3/library/zipfile.html

tensão, publicação de artigos ou registro de patentes/softwares. No módulo **Cálculo das produções** foram gerados os *data frames* das produções organizadas por pesquisador.

Cabe ressaltar que apenas três docentes registraram em seus currículos as produções artísticas/culturais realizadas, visto que o campus apenas possui cursos da área de Agronomia, Computação, Física e Letras Português/Inglês, ou seja, esse tipo de atividade não é o tema dominante de pesquisa ou projeto da maioria dos docentes lotados em 2021 no campus, além de que alguns não atualizam constantemente seus currículos.

No módulo **Apresentação dos Resultados**, é exibida a taxa de produção científica do campus que no segundo trimestre de 2021 foi aproximadamente **56,56**. Os valores da Tabela 1 mostram as quantidades totais cada tipo de produção.

Produções bibliográficas		**Produções técnicas**		**Produções artísticas / culturais**	
Trabalhos em eventos	680	Software	25	Artes visuais	0
Artigos publicados	261	Patente	0		
Artigos aceitos para publicação	7	Apresentação de trabalho	593	Artes cênicas	26
Livros publicados ou organizados	10	Organização de evento	195		
Capítulos de livros publicados	130	Trabalho técnico	12	Música	0
Textos em jornais ou revistas	10	Curso de curta duração ministrado	109		

Tabela 1. Quantidades totais de cada tipo de produção

Por ser uma instituição de ensino e pesquisa, observa-se as maiores quantidades na coluna produções bibliográficas. Os trabalhos em eventos, artigos publicados e capítulos de livros publicados são as atividades predominantes do campus.

Sobre produções técnicas do campus, a Tabela 1 mostra que a principal realização técnica entre os pesquisadores é a apresentação de trabalho em evento como congressos ou encontros de iniciação científica, o que é um resultado comum dos materiais bibliográficos publicados.

A segunda realização técnica com maior quantidade é a organização de eventos. No campus, ocorrem anualmente eventos de palestras e minicursos organizados pelos docentes e discentes para áreas de Agronomia, Letras e Computação, além de eventos especiais ou mesas redondas de outras áreas. Ministração de cursos de curta duração é a terceira ação técnica com maior valor porque alguns docentes precisam ofertar cursos de extensão para completar suas cargas horárias. Ademais, alguns docentes organizam e ofertam minicursos nos eventos do campus.

Por fim, as atividades artísticas/culturais não são as produções dominantes do campus, como já foi relatado anteriormente. Artes cênicas e outras produções são as únicas encontradas nos currículos dos pesquisadores. Em vista disso, só foi possível implementar e testar essa funcionalidade de contagem porque foram utilizados currículos de outros campi e instituições com cursos na área de música ou artes cênicas ou artes visuais. Todavia, ainda é uma operação que precisa de mais testes de currículos.

6. Análise comparativa

O Quadro 1 é uma síntese comparativa entre os sistemas descritos na seção 3 com o protótipo desenvolvido neste trabalho.

Software	Como coleta os currículos?	Resolve o RECAPTCHA?	Há custo financeiro de extração?	Está atualizado?	Nível de Funcionalidades
Protótipo proposto*	Através dos grupos de pesquisas do campus, extrai o ID Lattes e baixa os currículos automaticamente.	SIM	SIM, $ 2,99 por 1000 RECAPTCHA.	SIM	INCOMPLETO
Sistema da UNESP	Por meio de listas com os endereços permanentes dos currículos dos docentes de cada unidade da UNESP, o sistema extrai o ID Lattes e baixa os currículos no formato XML.	NÃO	NÃO	NÃO	COMPLETO
lucyLattes	O usuário precisa baixar manualmente os currículos no formato XML.	NÃO	NÃO	SIM	COMPLETO
UFOP Ativa	Consegue automaticamente baixar e extrair dados do currículo lattes no formato XML por meio de um WebService SOAP.	NÃO PRECISA	NÃO	SIM	COMPLETO

Quadro 1. Análise comparativa dos extratores de dados do currículo lattes

Como pode ser observado nesse quadro, o único sistema equivalente que baixa automaticamente os currículos é o UFOP Ativa em razão de ter conexão direta com a Plataforma Lattes via Web Service, em outras palavras, não precisa resolver o CAPTCHA/reCAPTCHA. Essa conexão foi autorizada pela CNPq após solicitação assinada pelo reitor da universidade.

Para os demais sistemas revisados, o usuário precisa baixar manualmente os currículos ou no mínimo resolver o capctha para baixar automaticamente (sistema da UNESP). Uma lista dos endereços dos currículos ou de ID Lattes precisam ser fornecidos em todos os softwares relacionados. O protótipo proposto executa todos esses processos de forma automática, isto significa que o próprio protótipo identifica os docentes que serão analisados, extrai o ID Lattes de cada um e faz o *download* dos currículos. O usuário apenas precisa acionar a análise e todo o resultado é obtido dentro de 20 minutos.

O protótipo desse trabalho é considerado incompleto por não cumprir todos os objetivos definidos, porquanto ainda não está disponível numa versão final desktop para uso da coordenadora até este momento. Os demais sistemas são completos porque cumprem todos os objetivos definidos por seus autores, geram relatórios sumarizados do currículo e alguns até apontam inconsistências de informações. A segunda desvantagem deste protótipo é o custo financeiro para cada reCAPTCHA resolvido, apesar de ser baixo o valor anual, já que os demais sistemas não possuem custos financeiros para a análise de dados.

Por último, o presente protótipo com lucyLattes e UFOP Ativa são os únicos atualizados ou desenvolvidos recentemente que podem ser utilizados nessa tarefa de extração de dados do lattes.

7. Considerações finais

A produção de trabalhos da área de Ciências de Dados, mais especificamente, na extração de informações sobre a produção acadêmica brasileira pode incentivar novas pesquisas científicas ou contribuições técnicas. Outra contribuição deste artigo, é apontar ferramentas relacionadas e servir de base para criar muitos outros projetos relacionados sobre coleta e análise de dados podendo até mesmo aplicar técnicas de aprendizagem de máquina.

Baseado nos resultados relatados, este projeto mostrou ser uma alternativa eficaz para auxiliar coordenadores de pesquisas do IFCE no cálculo dos registros acadêmicos do seu campus. Ele é totalmente automático em suas funcionalidades de aquisição dos currículos, principal diferencial quando comparado com os sistemas revisados. O protótipo está passível de alterações que poderão ser futuramente aplicadas, como a implementação em versão *desktop*, a separação entre pesquisadores externos e internos do campus, a geração de relatórios mais completos sobre cada pesquisador e um *dashboard* com os seguintes indicadores: taxa de produção científica, um gráfico de barras da quantidade de cada tipo de realização por pesquisador podendo filtrar o tipo de atividade, um gráfico pizza com a porcentagem de ações totais de cada pesquisador em relação a somatória de todos.

Referências

Branco, A. M. et al. (2018). Ferramenta para coleta e comparação de dados de publicações acadêmicas dos professores com o currículo lattes. Monografia do Curso de Ciência da Computação da Universidade Federal de Santa Catarina (UFSC).

BRITO, A. G. C. d., Quoniam, L., and Mena-Chalco, J. P. (2016). Exploração da plataforma lattes por assunto: proposta de metodologia. *Transinformação*, 28(1):77–86.

Corrêa, T. S., Suzuki, M. B., Cintra, P. R., and Costa, L. S. F. (2017). O fim do scriptlattes? uma análise de suas funcionalidades, alternativas para o presente e perspectivas para o futuro. *Revista do EDICC-ISSN 2317-3815*, 3(3).

Lana, M. V. P. (2019). UFOP ativa: uma ferramenta para extração e análise de dados acadêmicos de servidores da UFOP baseada no currículo lattes. Monografia do curso de Ciência da Computação da Universidade Federal de Ouro Preto (UFOP).

Mena-Chalco, J. P. and Júnior, C. (2013). Prospecção de dados acadêmicos de currículos lattes através de scriptlattes. *Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João*, pages 109–128.

Plataforma Lattes (2021). Dgp: O que é. Disponível em `http://lattes.cnpq.br/web/dgp/o-que-e/`. Acesso em 22 de maio de 2021.

PRPI (2021). Plataforma NL da PRPI. Disponível em `http://prpi.ifce.edu.br/nl/app\_Login/`. Acesso em 02 de fevereiro de 2021.

Tieppo, R. (2019). lucyLattes um script para manipular dados da plataforma Lattes. Disponível em `https://rafatieppo.github.io/post/2019\_03\_13\_lucylattes/`. Acesso em 15 de maio de 2021.

Vidotti, S. A. B. G. (2015). Coleta de dados a partir dos currículos da Plataforma Lattes: procedimentos utilizados no repositório Institucional UNESP. *Ponto de Acesso*, 9(3):117–132.