ANÁLISE DE PRODUÇÃO CIENTÍFICA COM BASE EM CURRÍCULO ONLINE: UM PROTÓTIPO DE FERRAMENTA DE CÓDIGO ABERTO

Giovanna Cristina Martins [1], Nelson Nunes Tenório Junior [2]

[1]Giovanna Cristina Martins, Campus Maringá-PR, Universidade Cesumar - UNICESUMAR. Bolsista PIBIC/ICETI-UniCesumar.
Giovanna.cristina.acad@gmail.com

RESUMO

A Plataforma Lattes representa o principal repositório curricular da comunidade científica brasileira, reunindo dados relevantes sobre a trajetória acadêmica e produtiva de pesquisadores. Contudo, a extração e análise de informações em larga escala a partir desses currículos ainda apresenta barreiras técnicas e operacionais. Este projeto de iniciação científica propõe o desenvolvimento de um protótipo de ferramenta de código aberto para extração, organização e análise da produção científica disponível nos currículos Lattes. Utilizando tecnologias acessíveis e técnicas de mineração de dados, a proposta visa automatizar o processo de coleta de dados e gerar relatórios analíticos sobre a produção acadêmica. A ferramenta também permitirá a geração de mapas de colaborações e visualizações geográficas da atuação dos pesquisadores. Espera-se que os resultados contribuam para maior transparência, eficiência e reprodutibilidade na gestão de informações acadêmicas, democratizando o acesso a dados estratégicos da ciência brasileira.

PALAVRAS-CHAVE: Análise; Código livre; Extração de dados; GPL; Plataforma Lattes.

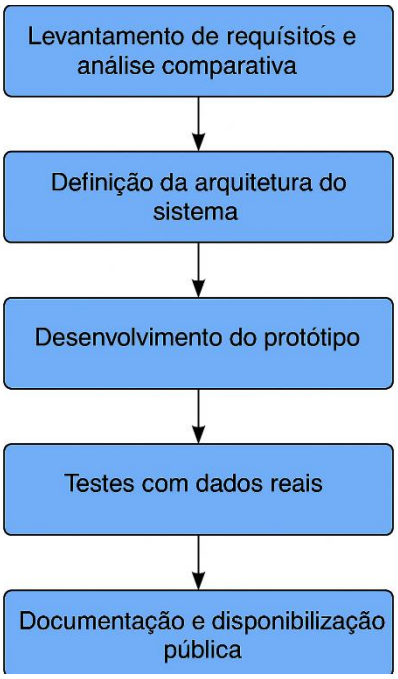
1 INTRODUÇÃO

A análise da produção científica tem ganhado destaque na gestão universitária e na formulação de políticas públicas voltadas à ciência, tecnologia e inovação. Nesse cenário, a Plataforma Lattes consolidou-se como a principal base de dados curricular da comunidade científica brasileira, reunindo informações detalhadas sobre formação acadêmica, publicações e vínculos institucionais de pesquisadores. O potencial informacional dessa base, no entanto, contrasta com as dificuldades operacionais para extração e tratamento automatizado dos dados, especialmente em larga escala.

2 MATERIAIS E MÉTODOS

Esta pesquisa caracteriza-se como aplicada e exploratória, estruturada em cinco etapas metodológicas interdependentes: (i) levantamento técnico e análise comparativa de ferramentas existentes; (ii) definição da arquitetura do sistema; (iii) desenvolvimento incremental do protótipo; (iv) testes com dados reais; e (v) documentação e disponibilização pública da solução, conforme a figura abaixo:

Quadro 1: etapas metodológicas interdependentes



O levantamento teórico será efetuado utilizando as seguintes palavras-chave em Língua Inglesa: “lattes” AND “tool” e, em Língua Portuguesa “lattes” AND “ferramenta”. As bases de dados utilizadas serão IEEExplore e ACM Digital Library, uma vez que tais bases são específicas para pesquisas que envolvem tecnologia. Além disso, buscar-se-á nessas bases artigos caracterizados como de acesso aberto, que não sejam de artigos de revisão e com um recorte temporal entre 2015 e 2025.

Na etapa inicial, será realizado um estudo comparativo de sistemas já consolidados na literatura, como o scriptLattes (MENA-CHALCO; JUNIOR, 2009), o LattesMiner (ALVES; YANASSE; SOMA, 2011b), o SUCUPIRA (ALVES; YANASSE; SOMA, 2011a) e o AnyLattes (CIRILO; SANTOS; MOTA, 2025) com o intuito de mapear suas funcionalidades, limitações e requisitos técnicos. Essa análise fundamentará o delineamento do escopo do protótipo a ser desenvolvido, permitindo alinhar as necessidades de usabilidade e reprodutibilidade com os avanços já alcançados na área de extração e análise de currículos acadêmicos.

A segunda etapa consistirá na concepção da arquitetura da ferramenta, baseada em princípios de modularidade, reutilização de código e baixo custo computacional. Será adotado o uso de linguagem Python pela sua ampla adoção em projetos científicos e disponibilidade de bibliotecas para web scraping (i.e., técnica utilizada pra a extração de dados de páginas web), análise de dados e visualização. Serão definidas as ferramentas e formas para parsing (i.e., verificação) da extração do HTML, Pandas para tratamento e estruturação dos dados, uma ferramenta de Business Intelligence para a geração de visualizações interativas dos gráficos de produção científica dos pesquisadores.

O desenvolvimento da ferramenta ocorrerá de modo iterativo, empregando princípios do desenvolvimento ágil (Scrum), com validações contínuas a cada módulo implementado. O sistema será dividido em quatro módulos principais: (1) um módulo de entrada, responsável pela leitura de listas de identificadores Lattes; (2) um módulo de extração, encarregado da coleta automatizada dos currículos em HTML; (3) um módulo de pré-processamento, voltado à normalização, remoção de duplicações e categorização das produções; e (4) um módulo analítico-visual, que fornecerá relatórios, gráficos temporais da produção científica.

Para validação funcional da ferramenta, será utilizada uma amostra real de currículos pertencentes a um programa de pós-graduação autorizado, a fim de aferir a acurácia dos dados extraídos e a coerência dos indicadores gerados. Essa validação será conduzida em conformidade com as diretrizes do CNPq sobre o uso público das informações contidas na Plataforma Lattes.

Por fim, a etapa conclusiva contemplará a documentação técnica do protótipo, incluindo instruções de uso, exemplos práticos, estrutura de entrada e saída, bem como os requisitos para instalação e execução do sistema. O código-fonte será disponibilizado em repositório público (por exemplo, GitHub), sob licença de software livre, com o objetivo de fomentar a replicabilidade do estudo e a colaboração contínua da comunidade científica na evolução da ferramenta.

3 RESULTADOS E DISCUSSÕES

Espera-se que, ao final da execução deste projeto, seja desenvolvido um protótipo funcional baseado em software de código aberto, capaz de automatizar a extração, organização e análise da produção científica registrada nos currículos da Plataforma Lattes.

A ferramenta deverá permitir a coleta estruturada de dados públicos a partir de listas de identificadores Lattes, viabilizando o tratamento e a padronização das informações extraídas, com foco na geração de indicadores de produção científica e representações visuais da atividade científica de pesquisadores. Adicionalmente, espera-se que o protótipo contribua diretamente para a sistematização de informações curriculares em processos de avaliação institucional, planejamento acadêmico e gestão da ciência. Além disso, são esperados desdobramentos acadêmicos e formativos relevantes. A ferramenta será devidamente documentada e publicada em repositório público, sob licença de software livre, garantindo sua ampla acessibilidade, transparência e possibilidade de reuso por outros pesquisadores e instituições. A experiência prática no desenvolvimento da do software deverá também resultar na redação e submissão de um artigo técnico ou científico, relatando a arquitetura do sistema, os métodos adotados e os resultados obtidos em seu uso experimental. Do ponto de vista formativo, o projeto propiciará ao bolsista o desenvolvimento de competências técnicas e científicas em ciência de dados, mineração de currículos acadêmicos e boas práticas de programação em ambientes de pesquisa. Espera-se, assim, a formação de um perfil discente interdisciplinar, apto a atuar na interface entre ciência e gestão da informação científica. Por fim, disponibilizando uma solução acessível, flexível e científica para análise de dados curriculares, o projeto contribuirá com a promoção da ciência aberta, ampliando as possibilidades de análise qualificada da produção científica brasileira, especialmente em contextos de avaliação e planejamento da pós-graduação.

4 CONSIDERAÇÕES FINAIS

Espera-se que, ao final da execução deste projeto, seja desenvolvido um protótipo funcional capaz de automatizar a extração e análise da produção científica registrada nos currículos da Plataforma Lattes. A disponibilização pública do código contribuirá para a democratização do acesso a dados estratégicos e para a promoção da ciência aberta, ampliando as possibilidades de avaliação institucional e colaborativa da produção acadêmica brasileira

REFERÊNCIAS

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks. In: 2011a, 6th Iberian Conference on Information Systems and Technologies (CISTI 2011). [S. l.: s. n.] p. 1–6.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. LattesMiner: a multilingual DSL for information extraction from lattes platform. *In*: 2011b, New York, NY, USA. **Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11**. New York, NY, USA: ACM, 2011. p. 85–92.Disponível em: https://doi.org/10.1145/2095050.2095065

CIRILO, A. C. S.; SANTOS, I. M. dos; MOTA, M. P. AnyLattes: An Application for Continuous Assessment of Lattes Curriculum Information. *In*: 2025, **Anais do XXI Simpósio Brasileiro de Sistemas de Informação (SBSI 2025)**. : Sociedade Brasileira de Computação, 2025. p. 439–448.Disponível em: https://doi.org/10.5753/sbsi.2025.246529

DIAS, T. M. R.; MOITA, G. F.; DIAS, P. M. Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na Plataforma Lattes. **Em Questão**, n. 1, p. 63–86, 2019. Disponível em: https://doi.org/10.19132/1808-5245251.83-86

FERRAZ, R. R. N.; QUONIAM, L. M.; MACCARI, E. A. A Utilização da Ferramenta ScriptLattes para Extração e Disponibilização Online da Produção Acadêmica de um Programa de Stricto Sensu em Administração. *In*: 2014, **Proceedings of the 11th CONTECSI International Conference on Information Systems and Technology Management**. : TECSI, 2014.Disponível em: https://doi.org/10.5748/9788599693100-11CONTECSI/PS-583

MATHEUS BOIJINK, F.; PRASS, F. S. **Extração de Dados Quantitativos do Currículo Lattes**. *[S. l.: s. n.]*.

MENA-CHALCO, J. P.; JUNIOR, R. M. C. 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. Disponível em: https://doi.org/10.1007/BF03194511