MineraLattes: Ferramenta Computacional para Análise Estruturada de Dados XML da Plataforma Lattes

Lucas de Almeida Fernandes[1]; 0000-0003-1695-6915
Italo Pinto Rodrigues[1]; 0000-0002-6832-8358

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
italoprodrigues@gmail.com (contato principal)

Resumo: A Plataforma Lattes constitui a principal base curricular da comunidade científica brasileira, reunindo informações essenciais para processos de avaliação e gestão acadêmica. Entretanto, a extração manual de dados a partir dos currículos é trabalhosa e suscetível a erros, especialmente em levantamentos institucionais de grande porte. Este trabalho apresenta o desenvolvimento de um algoritmo em MATLAB para mineração de informações em arquivos XML extraídos da Plataforma Lattes. O código realiza a leitura estruturada dos currículos, identifica categorias de produção acadêmica, contabiliza ocorrências dentro de um intervalo temporal definido e exporta automaticamente planilhas organizadas em níveis de resumo e detalhamento. Nos testes realizados, os resultados obtidos coincidiram integralmente com os indicadores exibidos na própria plataforma, confirmando a precisão da abordagem. A proposta representa uma alternativa simples e eficiente para instituições que necessitam agilizar a coleta de dados acadêmicos, embora se reconheça a necessidade de testes adicionais para verificar sua robustez em diferentes contextos.

Palavras-chave: Plataforma Lattes. Mineração de dados. Automação. MATLAB. Produção científica.

INTRODUÇÃO

A Plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), consolidou-se como o principal sistema de registro e disponibilização de informações sobre a produção científica no Brasil. Nela, pesquisadores de diferentes áreas mantêm currículos padronizados, que são amplamente utilizados em processos de avaliação institucional, análise de produtividade acadêmica e elaboração de relatórios de desempenho (Cirilo; Santos; Mota, 2025; Mena-Chalco; Junior, 2013, 2009). Apesar da riqueza dos dados disponibilizados, a extração de informações em larga escala permanece um desafio, pois os registros estão estruturados em currículos individuais, exigindo consultas e compilações manuais, processo moroso e suscetível a erros.

Diante desse cenário, diversas ferramentas têm sido propostas para automatizar a mineração e a organização de dados do Lattes. Entre elas, destaca-se o scriptLattes, software livre pioneiro que permite a criação de relatórios acadêmicos a partir de currículos em formato HTML, incluindo análises bibliométricas, identificação de colaborações e geração de mapas de produção científica (Mena-Chalco; Junior, 2013, 2009). Mais recentemente, soluções como o AnyLattes ampliaram o escopo dessa automação ao incorporar a análise de indicadores exigidos pela CAPES e a integração com a Plataforma Sucupira, otimizando a rotina de avaliação de programas de pós-graduação (Cirilo; Santos; Mota, 2025).

Apesar dos avanços, tais ferramentas ainda apresentam limitações importantes. O scriptLattes, por exemplo, opera a partir de currículos em HTML, exigindo adaptações complexas para lidar com dados em XML, formato disponibilizado diretamente pela Plataforma Lattes (Mena-Chalco; Junior, 2013, 2009). Já o AnyLattes, embora ofereça integração avançada com métricas de avaliação, requer configurações específicas em ambientes web, o que pode restringir sua adoção em contextos institucionais mais simples (Cirilo; Santos; Mota, 2025). Nesse contexto, observa-se uma lacuna na disponibilidade de soluções leves, multiplataforma e de fácil implementação, capazes de explorar diretamente arquivos XML fornecidos pelo Lattes e de gerar relatórios prontos para análise.

Com o objetivo de preencher essa lacuna, este trabalho apresenta o desenvolvimento de um código em MATLAB voltado à mineração de informações em arquivos XML extraídos da Plataforma Lattes. A solução proposta realiza a leitura estruturada dos currículos, identifica

categorias de produção acadêmica e gera automaticamente planilhas organizadas, permitindo o levantamento de dados de pesquisadores de uma instituição de forma ágil e confiável.

MÉTODOS

Este estudo baseou-se no desenvolvimento de um algoritmo em MATLAB voltado à mineração de informações em arquivos XML da Plataforma Lattes. O código foi projetado de forma modular, estruturando-se em etapas encadeadas que vão da seleção de arquivos à exportação final dos resultados.

Na primeira etapa, o usuário seleciona um ou mais arquivos XML de interesse. O sistema verifica automaticamente a integridade e compatibilidade desses arquivos, interrompendo a execução em caso de inconsistências. Em seguida, solicita-se a definição de um intervalo de anos, de modo a restringir a análise ao período especificado, prática essencial para relatórios institucionais anuais ou quadrienais.

A etapa seguinte corresponde ao mapeamento de categorias acadêmicas às marcações do XML. Foram consideradas produções bibliográficas, técnicas e artísticas, além de atividades como orientações, bancas e eventos. Cada categoria foi associada a um conjunto de tags do Lattes, o que permite a extração automática dos registros sem intervenção manual.

Uma vez definido o mapeamento, o algoritmo realiza a extração e contagem, percorrendo os arquivos para identificar as tags correspondentes e contabilizar ocorrências por categoria. Para cada item, verifica-se o ano de publicação ou defesa, assegurando que apenas dados dentro do intervalo informado sejam considerados. O sistema mantém acumuladores por categoria e ano, possibilitando a geração de estatísticas agregadas e detalhadas.

Os resultados são então organizados em dois níveis complementares:

- Resumo por pesquisador, contendo o total de produções em cada categoria e um somatório geral;
- Detalhamento anual, estruturado em formato tabular, associando pesquisador, categoria, ano e número de itens.

Essas informações são exportadas automaticamente para uma planilha Excel, em abas distintas, organizadas de forma a facilitar consultas, comparações e uso posterior em softwares de análise.

O programa inclui ainda recursos de usabilidade e robustez. Entre eles, a geração de mensagens de erro em caso de intervalos inválidos ou falhas de leitura, além da possibilidade de execução em dois modos: normal, que apresenta apenas a mensagem de conclusão com o local do arquivo gerado, e de depuração, no qual são exibidos logs detalhados da execução, permitindo acompanhar o processamento em tempo real.

RESULTADOS E DISCUSSÃO

A aplicação do algoritmo em arquivos XML da Plataforma Lattes resultou na geração automática de planilhas organizadas em duas dimensões: um resumo por pesquisador, no qual constam os totais de produções discriminadas por categoria, e um detalhamento anual, que relaciona pesquisador, ano, categoria e número de itens. Essa estruturação permitiu uma visualização integrada da produção acadêmica, com possibilidade de análises longitudinais e comparativas.

Nos testes realizados, os resultados obtidos foram confrontados com os indicadores exibidos diretamente na Plataforma Lattes. Observou-se 100% de correspondência entre as contagens geradas pelo algoritmo e os registros oficiais, conforme comparação apresentada na Figura 1, evidenciando a precisão da abordagem empregada e a adequação do mapeamento de categorias às marcações do XML. Esse achado confirma a robustez do processo de mineração e valida a utilização do código como ferramenta confiável para levantamentos institucionais.

Figura 1 – Comparação entre os indicadores da Plataforma Lattes e os indicadores extraídos pelo MATLAB.

Categoria	Ano	Itens	Total
Apresentação de trabalho e palestra	2023	2	
Artigos completos publicados	2023	1	4
Artigos completos publicados	2024	1	
Artigos completos publicados	2025	2	
Bancas	2023	8	
Bancas	2024	7	
Bancas	2025	3	
Orientações e supervisões	2023	10	
Orientações e supervisões	2024	24	
Outras produções bibliográficas	2024	1	1
Outras produções técnicas	2024	6	
Trabalhos publicados em anais de eventos	2023	13	35
Trabalhos publicados em anais de eventos	2024	15	
Trabalhos publicados em anais de eventos	2025	7	
Trabalhos técnicos	2023	6	
Trabalhos técnicos	2024	3	
Trabalhos técnicos	2025	5	

Fonte: Elaborada pelos autores (2025).

A automatização proporcionou ainda ganhos significativos em relação ao levantamento manual, uma vez que eliminou a necessidade de compilações individuais e reduziu o tempo de processamento de forma substancial. Além disso, a utilização de categorias pré-mapeadas assegurou consistência na contabilização, evitando variações decorrentes de interpretações humanas.

Comparado a iniciativas prévias, como o scriptLattes, o presente código diferencia-se por operar diretamente sobre os arquivos XML, dispensando a conversão para HTML e ampliando a compatibilidade com os formatos atualmente disponibilizados pela Plataforma Lattes (Mena-Chalco; Junior, 2013, 2009). Em relação ao AnyLattes, que privilegia análises voltadas a indicadores de avaliação da CAPES, o algoritmo aqui desenvolvido apresenta simplicidade de uso e maior portabilidade, sendo executável localmente sem a necessidade de configurações em ambiente web (Cirilo; Santos; Mota, 2025).

Essa abordagem oferece uma alternativa mais acessível para instituições que demandam apenas a organização e o levantamento quantitativo de dados, mas não dispõem de infraestrutura ou conhecimento técnico para implantar soluções mais complexas. O uso do MATLAB como base também representa um diferencial, já que muitos grupos de pesquisa e universidades já utilizam esse ambiente em rotinas científicas, facilitando a adoção.

CONCLUSÕES

O presente trabalho apresentou o desenvolvimento de um algoritmo em MATLAB voltado à mineração de informações a partir de arquivos XML da Plataforma Lattes. A solução implementada demonstrou-se eficaz na extração estruturada de dados, organizando a produção acadêmica em planilhas que contemplam tanto o resumo por pesquisador quanto o detalhamento anual por categoria.

A principal contribuição do estudo consiste em oferecer uma alternativa simples, precisa e de fácil implementação para instituições que necessitam levantar informações de forma ágil, reduzindo o tempo de análise e eliminando etapas manuais suscetíveis a erros. Nos testes realizados, verificou-se total correspondência com os indicadores fornecidos pela própria Plataforma Lattes, evidenciando a consistência do procedimento.

Apesar dos resultados positivos, destaca-se a necessidade de testes exaustivos em diferentes conjuntos de dados e contextos institucionais para confirmar a robustez do algoritmo e ampliar sua aplicabilidade. Estudos futuros podem explorar a integração da ferramenta com sistemas de gestão acadêmica, bem como a adaptação do código para outros ambientes de programação ou linguagens de uso aberto.

AGRADECIMENTOS

Os autores agradecem ao Centro Universitário de Volta Redonda (UniFOA) pelo apoio institucional e financeiro, por meio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) (92847/17/RPE).

REFERÊNCIAS

CIRILO, Alex Carlos Saraiva; SANTOS, Isadora Mendes Dos; MOTA, Marcelle Pereira. AnyLattes: An Application for Continuous Assessment of Lattes Curriculum Information. *In*: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 19 maio 2025. **Anais do XXI Simpósio Brasileiro de Sistemas de Informação (SBSI 2025)** [...]. Brasil: Sociedade Brasileira de Computação, 19 maio 2025. p. 439–448. DOI 10.5753/sbsi.2025.246529. Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/34360. Acesso em: 16 set. 2025.

MENA-CHALCO, Jesús Pascual; JUNIOR, Roberto Marcondes Cesar. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. [*s. l.*], 2013. DOI

10.13140/RG.2.1.5183.8561. Disponível em: http://rgdoi.net/10.13140/RG.2.1.5183.8561. Acesso em: 16 set. 2025.

MENA-CHALCO, Jesús Pascual; JUNIOR, Roberto Marcondes Cesar. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, [*s. l.*], v. 15, n. 4, p. 31–39, dez. 2009. https://doi.org/10.1007/BF03194511.