



10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



LATTES IA:

uma plataforma aberta para análise e exploração semântica de currículos da Plataforma Lattes

Thiago Magela Rodrigues Dias

<https://orcid.org/0000-0001-5057-9936>

thiogomagela@cefetmg.br

Centro Federal de Educação Tecnológica
de Minas Gerais (CEFET-MG)

Heitor Rodrigues Carvalho

<https://orcid.org/0009-0001-2944-2094>

heitorrodrigues2909@gmail.com

Centro Federal de Educação Tecnológica
de Minas Gerais (CEFET-MG)

Washington L. R. Carvalho Segundo

<https://orcid.org/0000-0003-3635-9384>

washingtonsegundo@ibict.br

Instituto Brasileiro de Informação em
Ciência e Tecnologia (Ibict)

Patrícia Mascarenhas

<https://orcid.org/0000-0002-8448-6874>

patriciamdias@gmail.com

Universidade do Estado de Minas Gerais
(UEMG)

Resumo:

Este trabalho apresenta a plataforma Lattes IA, uma ferramenta aberta desenvolvida para análise de grandes conjuntos de currículos da Plataforma Lattes. A solução integra técnicas de mineração de dados, visualização interativa e inteligência artificial para explorar dados acadêmicos em larga escala. O sistema permite o carregamento de conjuntos institucionais de currículos, a geração automática de dashboards analíticos e a interação por meio de agentes conversacionais capazes de responder consultas semânticas sobre produção científica e trajetórias acadêmicas. A proposta amplia as possibilidades de exploração de dados curriculares e contribui para o desenvolvimento de infraestruturas abertas de análise da ciência brasileira.

Palavras-Chave: Inteligência artificial. Mineração de dados científicos. Ciência aberta.

EIXO TEMÁTICO:

Técnicas, Ferramentas e
Infraestruturas

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.1150

COMO CITAR ESTE ARTIGO:

DIAS, Thiago Magela Rodrigues; CARVALHO, Heitor Rodrigues; MASCARENHAS, Patrícia; CARVALHO SEGUNDO, Washington L. R. Lattes IA: uma plataforma aberta para análise e exploração semântica de currículos da Plataforma Lattes. *In*: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1150. Disponível em: <https://doi.org/10.22477/x.ebbc.1150>.



1 INTRODUÇÃO

A crescente disponibilidade de dados científicos estruturados tem ampliado significativamente as possibilidades de análise da atividade acadêmica em larga escala. No contexto brasileiro, a Plataforma Lattes constitui uma das mais abrangentes infraestruturas de informação científica, reunindo milhões de currículos que registram trajetórias acadêmicas, vínculos institucionais, produção científica e participação em atividades de pesquisa. Essa base de dados tornou-se uma fonte estratégica para estudos bibliométricos e cientométricos, bem como para avaliações institucionais e políticas públicas de ciência e tecnologia (Lane, 2010).

Apesar da riqueza informacional desses dados, a exploração analítica de grandes conjuntos de currículos ainda apresenta desafios técnicos relevantes. O volume e a complexidade estrutural dos arquivos XML, combinados com a heterogeneidade das informações registradas pelos pesquisadores, tornam o processamento e a análise desses dados uma tarefa computacionalmente complexa. Além disso, a maior parte das ferramentas disponíveis para análise da Plataforma Lattes exige conhecimentos técnicos em programação e manipulação de bases de dados, o que limita o acesso de pesquisadores e gestores institucionais a essas análises. Atualmente diversas ferramentas se propõem a analisar currículos da Plataforma Lattes (Mena-Chalco; Cesar Junior, 2013; Almeida Fernandes; Rodrigues, 2025; Segundo et al., 2025; Dias; Segundo; Souza, 2025).

Nos últimos anos, o avanço das técnicas de mineração de dados, visualização interativa e modelos de linguagem de grande escala (*Large Language Models – LLMs*) tem possibilitado novas formas de exploração de grandes bases de dados científicos. Em particular, abordagens baseadas em geração aumentada por recuperação (*Retrieval-Augmented Generation – RAG*) permitem combinar bases estrutura-

das de conhecimento com capacidades avançadas de interpretação semântica, possibilitando consultas em linguagem natural e geração automática de sínteses analíticas.

Nesse contexto, o objetivo deste trabalho é apresentar o desenvolvimento da plataforma Lattes IA, uma infraestrutura aberta para análise e exploração de dados curriculares da Plataforma Lattes, bem como discutir suas capacidades de integração entre processamento de dados, visualização interativa e interação semântica baseada em inteligência artificial.

Dessa forma, a proposta busca contribuir para o desenvolvimento de novas infraestruturas abertas de análise da ciência brasileira, ampliando as possibilidades de exploração e interpretação de grandes conjuntos de dados curriculares acadêmicos.

Para além de suas capacidades técnicas, a plataforma proposta apresenta implicações diretas para a gestão da informação científica em diferentes níveis institucionais. Em programas de pós-graduação, por exemplo, a ferramenta pode apoiar o monitoramento da produção científica de docentes e discentes, contribuindo para processos de avaliação e planejamento estratégico, especialmente em contextos como a avaliação da pós-graduação conduzida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

2 METODOLOGIA

O processo de coleta e análise dos dados foi estruturado para permitir o processamento eficiente de grandes volumes de dados curriculares acadêmicos, garantindo modularidade, escalabilidade e rastreabilidade. O fluxo metodológico é organizado em três macro etapas: extração e transformação de dados, processa-

Figura 1 – Pipeline da Plataforma Lattes IA.

O pipeline é dividido em três etapas principais:

- Extração e Tratamento dos Dados:** A Plataforma Lattes (CNPq) é utilizada para o Download dos currículos. Esses dados são convertidos de XML para JSON por um Script de conversão de XML para JSON. O Banco de dados JSON é então processado por um Script de geração de tabelas CSV, resultando no Banco de dados CSV. Outro Script de geração de texto Markdown processa o Banco de dados JSON para gerar o Banco de dados Markdown.
- Processamento:** Os dados são lidos e processados usando pandas. O Banco de dados CSV é usado para alimentar o Modelo de Embeddings, que gera a Indexação vetorial. O Banco de dados Markdown é usado para alimentar o Modelo de Embeddings, que gera o Banco de dados vetorial. O Banco de dados XML é usado para alimentar o Modelo de Embeddings, que gera o Banco de dados vetorial. O Banco de dados vetorial é usado para alimentar o Agente de IA.
- Visualização:** O Agente de IA gera a Interface e plotagem de dados com React, que produz Gráficos e tabelas. O Agente de IA também responde a Perguntas do usuário através de um Chatbot, gerando a Resposta gerada.

O *pipeline* metodológico completo da plataforma proposta apresentado, evidencia as etapas de transformação dos dados curriculares em diferentes camadas de processamento analítico. Inicialmente ocorre a ingestão dos arquivos XML da Plataforma Lattes, seguida pela conversão para estruturas JSON que permitem manipulação programática mais eficiente. Posteriormente, os dados passam por processos de normalização e armazenamento, sendo então utilizados tanto para análises estruturadas quanto para a geração de representações vetoriais utilizadas em mecanismos de busca semântica e interação com modelos de linguagem.

10º EBBC, Curitiba - PR, Jul. 2026

quicas e maior compatibilidade com bibliotecas de análise de dados.

Após a transformação inicial, os dados passaram por etapas de limpeza, normalização e padronização, incluindo tratamento de valores ausentes e inconsistentes. Todo o conjunto de dados estruturados e textuais foi então armazenado em um banco de dados PostgreSQL, estabelecendo uma arquitetura de persistência unificada capaz de suportar consultas analíticas e recuperação semântica.

Os dados estruturados alimentam análises quantitativas que permitem examinar padrões acadêmicos, como distribuição temporal de publicações, frequência de coautoria e relações institucionais. Esses indicadores são apresentados por meio de visualizações interativas que permitem explorar tendências e correlações em diferentes níveis de agregação.

Os conteúdos textuais dos currículos foram processados para geração de *embeddings* vetoriais utilizando o modelo *multilíngue Qwen3-Embedding-0.6B*. Os vetores resultantes foram armazenados e indexados no PostgreSQL por meio da extensão *pgvector*, permitindo buscas eficientes por similaridade semântica.

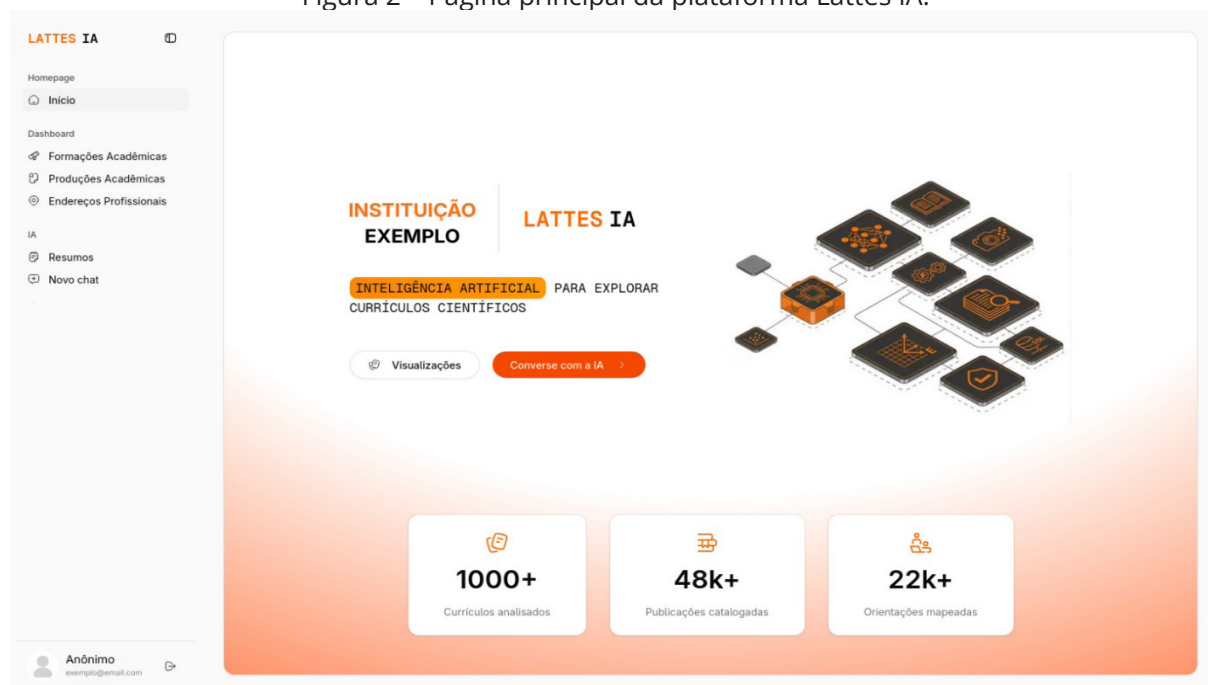
Sobre essa infraestrutura foi implementado um sistema de geração aumentada por recuperação de caráter agêntico, no qual o modelo de IA pode decidir autonomamente quais consultas realizar ao banco de dados e quantas iterações executar antes de formular uma resposta. O sistema também inclui geração automática de resumos acadêmicos a partir das informações do currículo Lattes. Em ambas as funcionalidades, o modelo de linguagem utilizado é baseado na API do Google AI Studio, por meio dos modelos Gemini, que fornecem suporte à capacidade de raciocínio e geração textual do sistema. A interação ocorre por meio de uma aplicação web desenvolvida em [Next.js](#), que integra visualizações de dados e uma interface conversacional baseada em IA.

3 RESULTADOS

O sistema implementado foi testado com um conjunto de dados representativo extraído da Plataforma Lattes pelo mecanismo de convênio institucional, compreendendo aproximadamente 2.6 mil perfis de pesquisadores. Os resultados demonstram que a integração de análises estruturadas e pesquisa conversacional baseada em RAG fornece uma abordagem complementar e sinérgica para explorar dados acadêmicos.

A interface da ferramenta foi dividida em três seções principais. A primeira se trata da página de início. Essa tela é responsável por apresentar a solução ao usuário, dados básicos do conjunto em análise e indicar suas principais funcionalidades (Figura 2).

Figura 2 – Página principal da plataforma Lattes IA.



Fonte: Os autores (2026).

O segundo módulo refere-se à visualização das análises dos currículos científicos. Essa seção foi projetada para apresentar os resultados de forma exploratória por

meio de múltiplas visualizações interativas, como por exemplo:

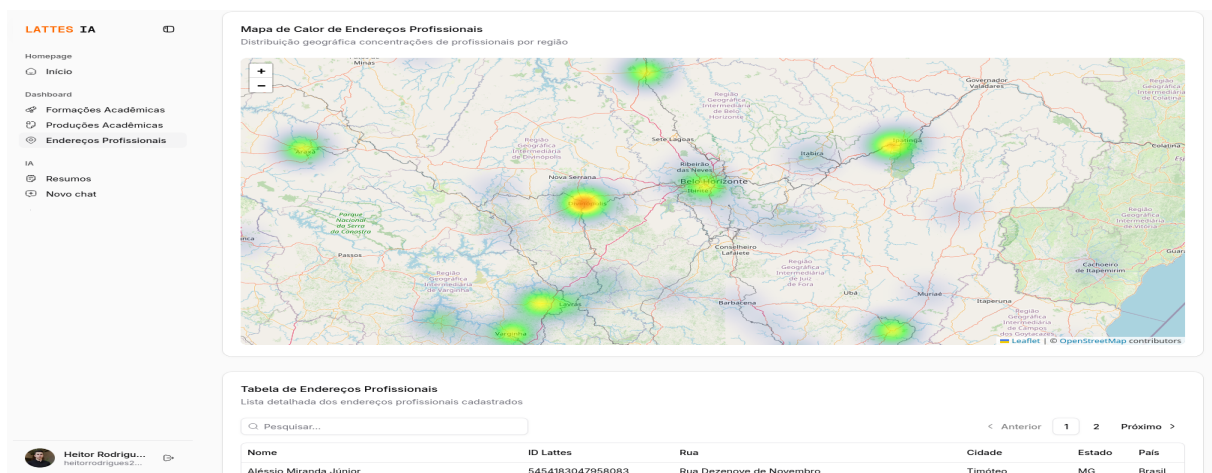
Tendências temporais: nessa interface, os usuários podem observar o volume de publicações por meio de gráficos interativos que permitem acompanhar a evolução das pesquisas ao longo dos anos, identificando períodos de crescimento ou estagnação em determinadas áreas.

Análises estruturadas: complementarmente, tabelas possibilitam a consulta detalhada dos dados agregados, facilitando comparações e análises de produtividade científica, através de sistemas de filtragem sofisticados.

Análise de palavras-chave: as frequências normalizadas de palavras-chave auxiliam a evidenciar temas emergentes dentro do conjunto de publicações analisado.

Comparação regional: essa interface permite visualizar padrões de intensidade ou concentração nos dados, contribuindo para a identificação de relações entre diferentes entidades distribuídas geograficamente (Figura 3).

Figura 3 – *Dashboard* de visualização dos dados.

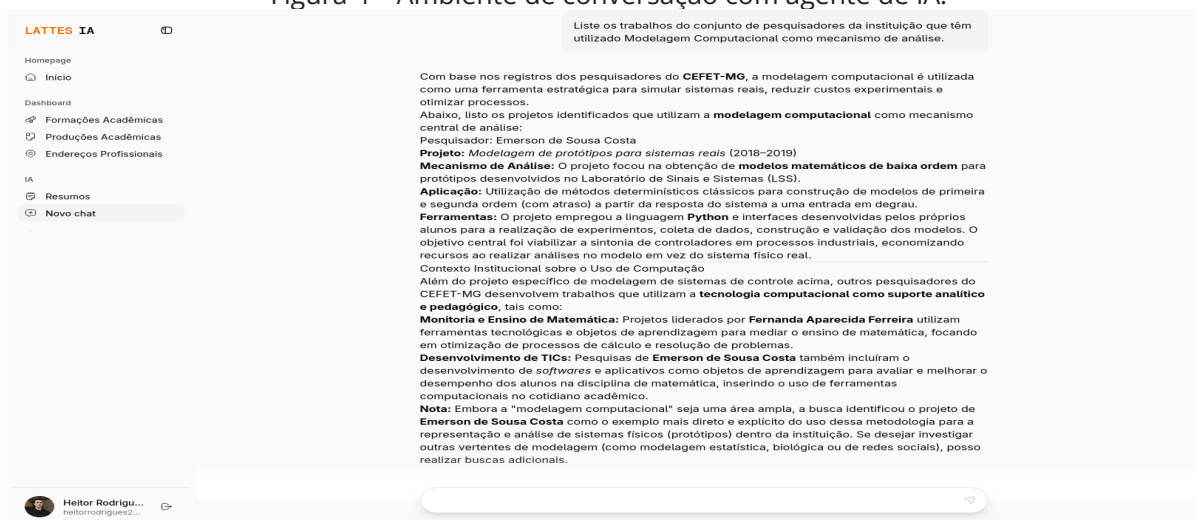


Fonte: Os autores (2026).

O terceiro segmento concentra as funcionalidades impulsionadas por LLMs. Nes-

sa interface, é disponibilizado um *chatbot* baseado em recuperação aumentada por geração (RAG), capaz de responder consultas em linguagem natural utilizando dados estruturados e textuais extraídos dos currículos analisados. Esse sistema permite que os usuários realizem perguntas sobre produção científica, trajetórias acadêmicas e relações de pesquisa, obtendo respostas contextualizadas a partir das fontes disponíveis. Além disso, a seção inclui um redator de resumos acadêmicos, desenvolvido para sintetizar automaticamente informações dos currículos, produzindo descrições concisas das atividades e contribuições dos pesquisadores (Figura 4).

Figura 4 – Ambiente de conversação com agente de IA.



Fonte: Os autores (2026).

De forma geral, os resultados demonstram que a plataforma proposta possibilita uma exploração mais abrangente e acessível de grandes conjuntos de dados curriculares provenientes da Plataforma Lattes. A integração entre visualizações interativas, análises estruturadas e mecanismos de interação baseados em inteligência artificial amplia significativamente as possibilidades de interpretação dos dados científicos, permitindo que usuários identifiquem padrões institucionais, tendências de produção científica e relações entre pesquisadores de maneira mais intuitiva.

4 CONSIDERAÇÕES

Os resultados demonstram que a integração entre análises estruturadas e mecanismos de interação semântica baseados em modelos de linguagem amplia significativamente as possibilidades de exploração dos dados científicos. Enquanto os *dashboards* interativos permitem identificar padrões quantitativos relacionados à produção acadêmica e às relações institucionais, o sistema conversacional baseado em recuperação aumentada possibilita consultas em linguagem natural e geração automática de sínteses informacionais a partir do conjunto de currículos analisados.

Essa abordagem evidencia o potencial das tecnologias de inteligência artificial como suporte à análise de grandes bases de dados científicos, contribuindo para novas formas de descoberta e interpretação da informação acadêmica. Além disso, ao permitir o carregamento de conjuntos específicos de currículos institucionais, a ferramenta pode apoiar processos de análise institucional, planejamento estratégico e avaliação científica.

REFERÊNCIAS

ALMEIDA FERNANDES, Lucas; RODRIGUES, Italo Pinto. MineraLattes: ferramenta computacional para análise estruturada de dados XML da plataforma Lattes. In: **Tudo é Ciência: Congresso Brasileiro de Ciências e Saberes Multidisciplinares**, [S. l.], n. 4, 2025. DOI: <https://doi.org/10.47385/tudoeciencia.2445.2025>. Disponível em: <https://conferencias.unifoa.edu.br/tc/article/view/2445>. Acesso em: 20 jun. 2026.

DIAS, Thiago Magela Rodrigues; SEGUNDO, Washington; SOUZA, Marcel Garcia. BrCris: facilitando o caminho para a Ciência Aberta através da exportação de conjuntos de dados. *Ciência da Informação*, v. 54, n. 2, 2025. DOI: <https://doi.org/10.18225/ci.inf.v54i2.7205>. Disponível em: <https://revista.ibict.br/ciinf/article/view/7205>. Acesso em: 20 jun. 2026.

LANE, Julia. Let's make science metrics more scientific. **Nature**, [S. l.], v. 464, n.



7288, p. 488-489, 2010. [Opinion]. DOI: <https://DOI.org/10.1038/464488a>. Disponível em: <https://www.nature.com/articles/464488a>. Acesso em: 15 fev. 2026.

MENA-CHALCO, Jesus P.; CESAR JÚNIOR, Roberto Marcondes. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. In: HAYASHI, Maria Cristina Piumbato Innocentini; LETA, Jacqueline. **Bibliometria e Cientometria: reflexões teóricas e interfaces**. São Carlos: Pedro & João Editores, 2013. p. 109-128. DOI: <https://doi.org/10.13140/RG.2.1.5183.8561>.

SEGUNDO, Washington Luís Ribeiro de; SOUZA, Marcel Garcia de; DIAS, Thiago Magela Rodrigues; CANTO, Fabio Lorensi do; SENA, Priscila Machado Borges. Dados científicos confiáveis na ciência aberta: desafios e estratégias de gestão e governança no BrCris. **Conferencia Internacional BIREDIAL-ISTEC**, [S. l.], 2025. DOI: <https://doi.org/10.22477/xiv.biredial.405>. Disponível em: <https://submissions.istec.org/index.php/biredial-istec/article/view/405>. Acesso em: 20 jun. 2026.