



UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº 17.228 de 25/11/2016



PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIX SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS **SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2025**

Ferramenta computacional para análise de currículos e colaborações em **Humanidades Digitais**

Paulo Henrique Dantas Barreto¹; Angelo Conrado Loula²

1. Paulo Henrique Dantas Barreto, PIBIC/FAPESB, ENGENHARIA DE COMPUTAÇÃO, e-mail: pdantasbarreto@protonmail.ch

2. Angelo Conrado Loula, DEPARTAMENTO DE CIÊNCIAS EXATAS, e-mail: angelocl@uefs.br

PALAVRAS-CHAVE: Humanidades Digitais; Colaboração Científica; Lattes

INTRODUÇÃO

As Humanidades Digitais (HD) vêm se consolidando, nas últimas décadas, como um campo interdisciplinar de conhecimento estruturado pela criação de associações, centros de pesquisa, programas acadêmicos e eventos científicos. Esse domínio integra as ciências humanas, as ciências da informação e a computação, propondo uma articulação entre os métodos tradicionais das humanidades e as potencialidades das tecnologias da informação (TI). Mais do que a simples digitalização de conteúdos analógicos, as HD configuram um novo paradigma epistemológico e metodológico, que redefine a maneira como o conhecimento é produzido, disseminado e analisado no ambiente digital (Guerreiro, 2015).

Dentro desse contexto, destaca-se a análise de redes de colaboração científica, uma vertente de investigação que busca compreender as interações entre pesquisadores e suas produções acadêmicas. A colaboração científica constitui um meio fundamental para aprimorar a qualidade e o impacto das pesquisas, uma vez que promove a integração de perspectivas diversas e a ampliação da troca de informações entre grupos e instituições (Balancieri et al., 2005). A compreensão dessas redes é essencial para a formulação de políticas de incentivo à pesquisa, otimização de recursos e fortalecimento de comunidades científicas.

Para apoiar esse tipo de análise, foram desenvolvidas ferramentas específicas que permitem identificar e visualizar relações de coautoria a partir de dados disponíveis na Plataforma Lattes — um repositório mantido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), órgão vinculado ao Ministério da Ciência e Inovação do Brasil (Brasil, 2022). Entre as soluções existentes, destacam-se o LucyLattes (<https://rafatieppo.github.io/lucylattes/>) e o ScriptLattes (<http://scriptlattes.sourceforge.net/>), que possibilitam a extração, organização e visualização textual ou gráfica de currículos de pesquisadores registrados na referida plataforma.

Neste trabalho, propõe-se o desenvolvimento de uma ferramenta computacional voltada à filtragem e indexação de currículos de pesquisadores cadastrados na Plataforma Lattes. Essa ferramenta tem como objetivo principal criar um índice que

permita a consulta e a seleção de currículos com base em múltiplos critérios definidos a partir de metadados estruturados. O processo de tratamento dos dados seguirá o modelo *Extract, Transform and Load* (ETL), conforme definido pela Amazon Web Service (AWS, 2025), visando à consolidação eficiente das informações em um banco de dados relacional.

A partir dessa filtragem, será possível gerar listas de currículos que atendam a critérios específicos, facilitando a extração e a análise de dados relevantes para estudos sobre produção científica e redes de colaboração. Em uma segunda etapa, prevista para ser desenvolvida em continuidade por outro pesquisador, a ferramenta será aprimorada para realizar a extração detalhada de informações sobre produções acadêmicas, projetos e orientações, incorporando também funcionalidades de análise estatística descritiva e modelagem das colaborações científicas.

MATERIAL E MÉTODOS OU METODOLOGIA

A pesquisa dá continuidade a um projeto de iniciação científica anterior, que já havia realizado a análise da Plataforma Lattes e dos arquivos XML dos currículos, além da identificação de metadados relevantes e otimização inicial do código.

Nesta etapa, a metodologia envolveu em uma primeira etapa: a) Levantamento de novas informações relevantes para aprimorar a extração de dados; b) Finalização do pipeline ETL, com extração dos metadados, transformação e limpeza das informações, e carga em um banco de dados relacional; c) Implementação de mecanismos de recuperação, assegurando a integridade dos dados em caso de falhas; d) Otimização de desempenho do código e dos processos de execução; e) Aplicação *web*, integrando o *backend* de dados à interface de consulta e exportação. Com a interface de usuário, permitindo busca, filtragem e integração com a ferramenta Lucy Lattes.

Para a realização da metodologia, foram utilizadas a linguagem Python 3.12 juntamente ao Gerenciador de Banco de Dados Relacional PostgreSQL 17. Bibliotecas como FastAPI, Cyclopts, Jinja2, JinjaX auxiliaram na implementação da interface Web e interface por linha de comando. ChartDB, foi utilizado para a modelagem do banco de dados, Beekeeper Studio para análise de buscas nesse banco, já o Github foi utilizado como plataforma *host* para o repositório do código fonte do projeto.

RESULTADOS

Esta etapa do projeto concentrou-se na consolidação da ferramenta de extração e análise de currículos da Plataforma Lattes, aprimorando seu desempenho, estabilidade e integração com o banco de dados e a aplicação web. As modificações implementadas buscaram otimizar o processo de ingestão de dados, reduzir falhas durante o *parsing* dos arquivos XML e oferecer uma interface mais acessível para o gerenciamento e consulta das informações coletadas.

Devido à complexidade associada a *deadlocks*, optou-se por não utilizar *threads* no processamento. Em substituição, adotou-se uma abordagem determinística, permitindo a retomada segura do processamento de currículos em caso de falhas. Para preservar o desempenho, foi implementado o uso de *transactions* em *batch*, enviando

múltiplos currículos ao banco de dados em uma única operação. Em caso de erro, o sistema os registra em um *log*, possibilitando a retomada a partir do ponto de falha.

O design do banco de dados também passou por ajustes, em resposta às discussões realizadas em grupo sobre novas possibilidades e requisitos. O resultado final pode ser encontrado na **Figura 1**.

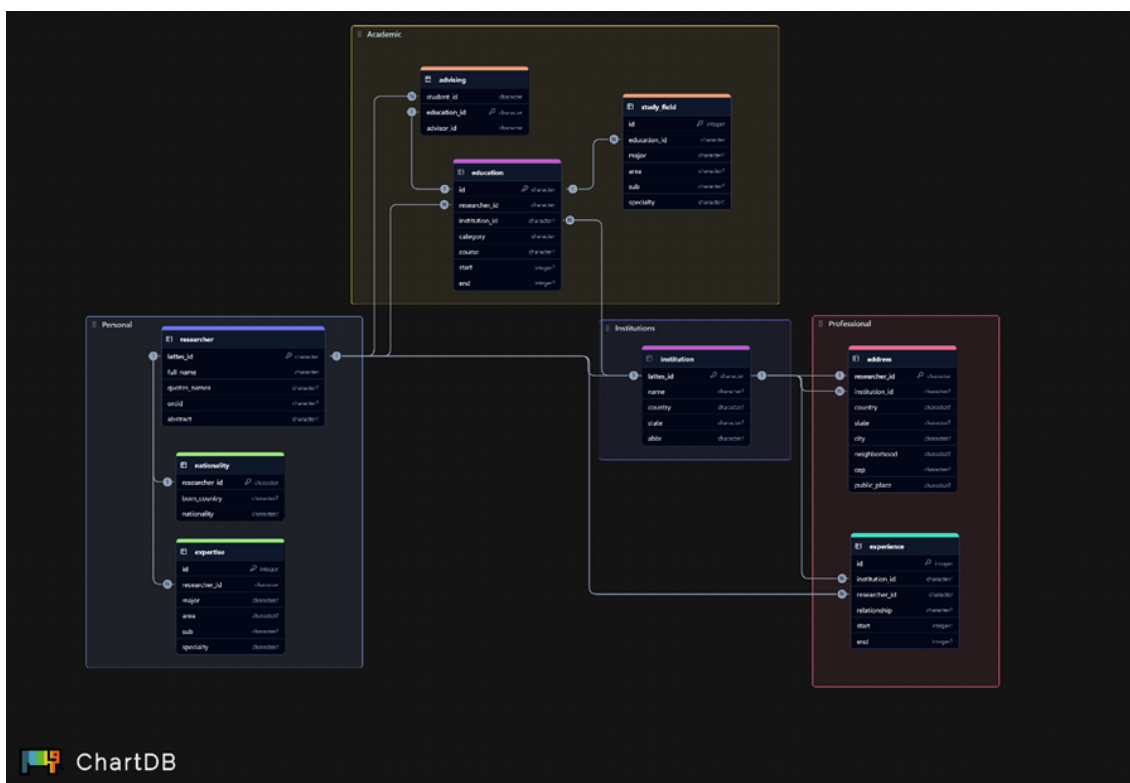


Figura 1: Modelagem final do banco de dados relacional do projeto. Na área Pessoal, temos as tabelas: Pesquisador, Nacionalidade e Especialidade; Na Acadêmica, temos Educação, Orientação e Campo de Estudo; Na Profissional, o Endereço Profissional e Experiência; E uma tabela especial de Instituição.

As tabelas foram modeladas com base na estrutura dos arquivos XML da Plataforma Lattes, de modo a preservar ao máximo a integridade e a completude das informações. O processo de modelagem buscou evitar a perda de dados e a inserção de valores inválidos, garantindo consistência entre o formato original e o banco de dados relacional. A seleção dos dados relevantes para extração foi discutida ao longo do desenvolvimento, resultando na estrutura final apresentada.

Além disso, foi implementada uma interface de linha de comando (Command Line Interface, CLI) para o gerenciamento do projeto, com comandos específicos, como o de bootstrap, responsável por inicializar o repositório de forma padronizada, e o de ingestão, que executa o processo de ETL e realiza o registro dos currículos no banco de dados relacional. Posteriormente, foi adicionado um novo comando voltado à inicialização da aplicação web do projeto, denominada Vitae.

No que se refere à aplicação web, desenvolveu-se uma interface que permite a interação do usuário com o sistema, possibilitando a realização de pesquisas e filtros por pesquisadores e áreas de atuação, bem como a visualização de informações relevantes e a exportação de dados para o Lucy Lattes, conforme ilustrado na **Figura 2**.

Inicialmente construída com FastAPI e Jinja2 para renderização de templates, a aplicação evoluiu para o uso do JinjaX, um plugin que facilitou a manutenção e a organização do código frente ao aumento da complexidade do front-end.

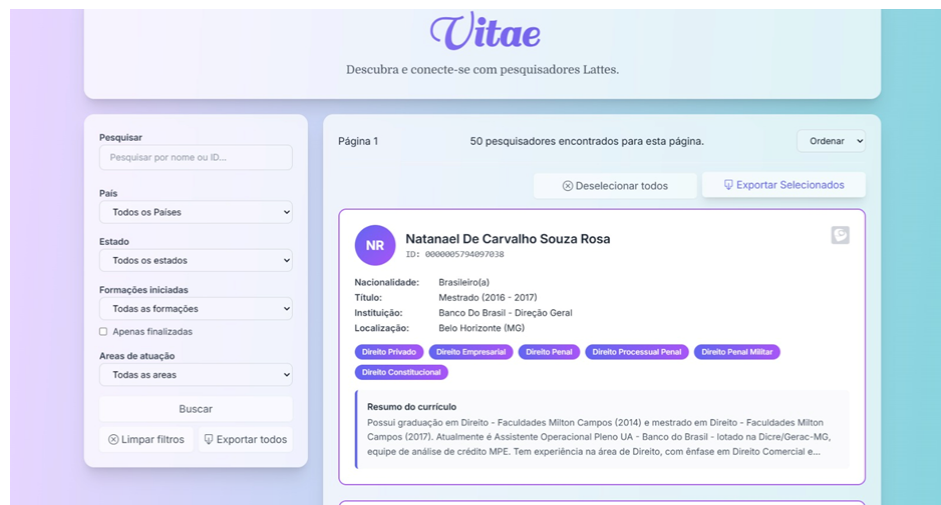


Figura 2: Visualização da aplicação Web denominada Vitae. Busca por todos os pesquisadores, sem filtros aplicados, com os pesquisadores selecionados para exportação Lucy Lattes em roxo.

CONSIDERAÇÕES FINAIS

O projeto consolidou uma ferramenta capaz de extrair e organizar dados dos currículos Lattes de forma confiável. A utilização de *transactions* em *batch* e da abordagem determinística garantiu a continuidade do processamento mesmo em caso de falhas, preservando a integridade dos dados. A modelagem do banco de dados, alinhada à estrutura dos arquivos XML, evitou perda ou inconsistência de informações, enquanto a interface de linha de comando e a aplicação web permitiram pesquisas, filtrações e exportações de dados de forma prática.

Dessa forma, a ferramenta oferece um pipeline robusto para o tratamento de informações acadêmicas, estabelecendo uma base sólida para futuras análises, integrações com novas fontes e a expansão das funcionalidades previstas, em linha com os objetivos inicialmente propostos.

REFERÊNCIAS

- BALANCIERI, R.; BOVO, A.B.; KERN, V.M.; PACHECO, R.C.D.S; BARCIA, R.M. 2005. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. *Ciência da informação*, 34, 64-77.
- GUERREIRO, D. 2014. *Humanidades Digitais: desafios e oportunidades*. Retrieved: https://dspace.uevora.pt/rdpc/bitstream/10174/12155/2/Dalia_Guerreiro_artigo_repositorio.pdf
- BRASIL. 2022. Ministério da Ciência e Inovação. CNPq - Plataforma Lattes (2022). Homepage: <https://lattes.cnpq.br>.
- Amazon Web Service (AWS). 2024. O que é ETL? – Explicação sobre extrair, transformar e carregar. Retrieved: <https://aws.amazon.com/pt/what-is/etl/>.