Universidade Federal de Minas Gerais

Instituto de Ciências Exatas – Departamento de Ciência da Computação

Mapeamento do Ecossistema de Inovação em Minas Gerais

Projeto Orientado em Computação II

Tipo de Pesquisa: **Tecnológica**

Orientador: **Marcos André Gonçalves**

Autor: Renato Silva Santos

Belo Horizonte – MG

2025

Resumo

Minas Gerais abriga um ecossistema de inovação vasto e diversificado com milhares de atores distribuídos por todo o estado. Entretanto, as informações sobre esses agentes encontram-se dispersas em múltiplas fontes, com diferentes níveis de atualização e padronização, o que dificulta o uso desses dados para análises, identificação de oportunidades de colaboração entre os atores e o desenvolvimento de ações de impacto. O Projeto Orientado em Computação I (POC I), desenvolvido no semestre anterior, concentrou-se na coleta, no tratamento e na classificação de dados de atores do ecossistema segundo as missões da política industrial Nova Indústria Brasil (NIB), a partir da base do Sistema Mineiro de Inovação (SIMI). No POC II, buscou-se a unificação dos dados heterogêneos de múltiplas fontes para a criação de uma plataforma integrada que permitisse explorar o ecossistema de inovação mapeado de forma unificada e estruturada. Para isso, consolidaram-se dados provenientes de quatro fontes distintas em um banco relacional, normalizando-os e organizando a caracterização dos atores segundo a perspectiva da hélice tríplice. Sobre essa base consolidada, aplicou-se uma nova classificação conforme as missões da Nova Indústria Brasil (NIB) e implementou-se um conjunto de serviços que deram origem à plataforma desenvolvida como parte do programa *Tem Base!* [1] — iniciativa do BH-TEC que visa conectar infraestrutura de pesquisa, setor produtivo e políticas públicas para impulsionar a inovação no estado. A entrega final compreende um dashboard interativo com painel para adição de novos registros e uma interface de *matchmaking* entre pesquisadores, constituindo uma ferramenta estratégica que viabiliza a identificação de competências, gargalos e oportunidades para acelerar a transferência de tecnologia e fortalecer a inovação em Minas Gerais.

1 Introdução

Minas Gerais abriga um ecossistema de inovação composto por empresas de base tecnológica, incubadoras, aceleradoras, instituições científicas e tecnológicas (ICTs) e outros atores estratégicos. Esses agentes atuam em segmentos variados — como saúde, agroindústria, tecnologias digitais, energia, bioeconomia, mobilidade e manufatura avançada — refletindo a riqueza e pluralidade do estado. No entanto, as informações sobre tais organizações encontram-se dispersas em múltiplas fontes, com diferentes padrões de registro, níveis de atualização e granularidade. Essa fragmentação dificulta uma visão integrada do ecossistema e limita o aproveitamento de oportunidades de colaboração, financiamento e transferência de tecnologia.

É neste contexto que origina-se esse projeto, realizado através do TCC Lab, um programa do BH-TEC. Especificamente, a iniciativa surgiu de um problema prático enfrentado pelo parque tecnológico na organização do *Matchday* — evento que aproxima pesquisadores, centros de tecnologia, empresas e indústria. A cada edição, o BH-TEC precisava identificar quais pesquisadores, laboratórios, centros tecnológicos e empresas que possuíam competências alinhadas ao tema definido, mas esse processo dependia majoritariamente de consultas manuais, redes de contato e recomendações informais. Tal abordagem era lenta, sujeita a lacunas e dependente do conhecimento dos envolvidos. Diante disso, tornou-se evidente a necessidade de uma ferramenta integrada que reunisse, estruturasse e disponibilizasse dados sobre o ecossistema de inovação mineiro.

À medida que o projeto avançava, evidenciava-se que essa infraestrutura de dados tinha potencial não apenas para apoiar o *Matchday*, mas também para orientar ações de médio e longo prazo do parque tecnológico, voltadas à articulação entre academia, indústria e governo. Paralelamente, o BH-TEC lançou o programa *Tem Base!* [1], iniciativa que propõe-se a conectar infraestrutura de pesquisa, setor produtivo e formulação de políticas públicas a partir de um mapeamento sistemático de laboratórios, centros de pesquisa e institutos de ciência e tecnologia em Minas Gerais. Como parte dessa iniciativa, foram coletados, através de formulários, dados sobre infraestruturas de pesquisa e unidades EMBRAPII atuantes no estado, classificados conforme as missões da Nova Indústria Brasil (NIB). Esses dados, juntamente com as bases já consolidadas no Sistema Mineiro de Inovação (SIMI), constituem a matéria-prima para este projeto.

Nesse contexto, o problema central enfrentado foi a ausência de uma plataforma integrada que consolide dados provenientes do SIMI, do mapeamento do *Tem Base!* e de outras possíveis fontes futuras, permitindo explorar de forma sistemática o ecossistema de inovação em Minas Gerais e sua relação com as missões da Nova Indústria Brasil sob a ótica da tríplice hélice. O projeto incorpora, além disso, o serviço de *matchmaking* desenvolvido no âmbito do *Tem Base!*, que, embora não tenha sido produzido por este trabalho, foi integrado à plataforma proposta, reunindo em uma única interface todos os esforços empreendidos para a consolidação do programa.

2 Fundamentação teórica

2.1 Ecossistemas de inovação e políticas orientadas por missão

A literatura sobre ecossistemas de inovação enfatiza a importância de arranjos institucionais que conectem universidades, empresas, governo e organizações intermediárias com objetivos compartilhados de desenvolvimento científico, tecnológico e econômico [2].

Nesse contexto, a política industrial Nova Indústria Brasil (NIB), lançada em 2024, oferece um enquadramento conceitual e operacional diretamente alinhado à realidade do projeto. A NIB constitui a estratégia federal de neoindustrialização do país, criada para reposicionar o Brasil em cadeias produtivas de maior valor agregado, fomentar a transição ecológica e digital, reduzir vulnerabilidades tecnológicas e ampliar a competitividade da indústria nacional. Estruturada em seis missões estratégicas, a política estabelece prioridades de investimento, inovação e desenvolvimento produtivo que orientam órgãos de fomento, agências de inovação e programas federais [**?**]. As seis missões da NIB são:

1. Descarbonização e transição energética para ampliar a sustentabilidade da base produtiva;
2. Transformação digital da indústria e fortalecimento de tecnologias críticas;
3. Saúde e complexo econômico-industrial da saúde para reduzir dependências externas;
4. Agroindústria sustentável e aumento da produtividade agrícola;
5. Mobilidade e logística sustentáveis, com foco em infraestrutura inteligente;
6. Infraestruturas, tecnologias e bens de capital para a soberania nacional.

Assim, classificar os atores do ecossistema mineiro segundo as missões da NIB equivale a produzir uma camada de informação imediatamente acionável para políticas públicas e outros múltiplos agentes. Adotar esse eixo de classificação temática representa um alinhamento ao mapeamento do ecossistema mineiro às prioridades nacionais. Isso viabiliza a análise sobre aderência das competências locais às missões, a identificação de lacunas e a oportunidades de reorientação de esforços de pesquisa, desenvolvimento e inovação. Portanto, a escolha da NIB, é justificada tanto pela sua centralidade na política brasileira atual quanto pela necessidade de tornar o mapeamento útil para a tomada de decisão dos *stakeholders*.

2.2 Tríplice hélice, setor institucional e tipo de organização

A perspectiva da tríplice hélice entende a inovação como resultado da interação dinâmica entre universidade, indústria e governo como esferas institucionais primárias [2]. No processo dessa interação surgem novas instituições secundárias, frequentemente denominadas "organizações híbridas", como parques tecnológicos, incubadoras, aceleradoras, associações empresariais e ambientes colaborativos, que atuam como interfaces entre os três vértices. Para que esse enquadramento teórico possa ser analisado empiricamente, é necessário traduzir essas esferas em categorias observáveis no conjunto de dados.

Neste trabalho, essa tradução é feita combinando duas dimensões de caracterização das organizações: o setor institucional ao qual pertencem e o tipo de papel que desempenham no ecossistema. Essas novas dimensões, criadas programaticamente, serão tratadas mais detalhadamente nas seções posteriores deste texto.

A classificação por setor correspondente à coluna `setor` da base de dados agrupa os atores em três blocos. O primeiro setor reúne as organizações públicas presentes na base, incluindo órgãos de governo, universidades e instituições públicas de ciência e tecnologia. O segundo setor é composto por empresas privadas e organizações cuja atuação segue

uma lógica predominantemente de mercado. O terceiro setor inclui organizações da sociedade civil, fundações e entidades sem fins lucrativos. Essa classificação dialoga com a lógica da Tríplice hélice, mas com um ajuste importante: governo e academia. Na formulação clássica ambos aparecem como esferas separadas, mas aqui são classificados conjuntamente por integrarem um agrupamento de organizações públicas do banco.

A segunda dimensão, denominada `tipo_organizacao`, diferencia as organizações segundo seu papel funcional no ecossistema. Essa coluna assume três valores: *Ambientes de inovação*, *Empresas* e *Infraestruturas de pesquisa*. Ambientes de inovação, que incluem parques tecnológicos, hubs, incubadoras, aceleradoras e centros de empreendedorismo, representam arranjos híbridos criados para estabelecer conexões entre governo, universidade, empresas e, em diversos casos, também o terceiro setor. Empresas correspondem à hélice indústria. Infraestruturas de pesquisa, como institutos, laboratórios e ICTs, aproximam-se da hélice academia ou ciência.

O cruzamento das colunas `setor` e `tipo_organizacao` na base integrada e sua disponibilização na plataforma de visualização permitem mapear empiricamente como os diferentes atores se distribuem entre as esferas da tríplice hélice. Esse procedimento possibilita identificar onde se concentram os ambientes híbridos de inovação e analisar de que forma governo, empresas, academia e sociedade civil se articulam em torno das missões da Nova Indústria Brasil.

3 Contribuições

O POC II partiu de uma base já estruturada do POC I, em que foram conduzidos o tratamento, a normalização e uma primeira camada de classificação das organizações do SIMI segundo as missões da Nova Indústria Brasil. Nesta etapa, o foco deslocou-se para: (i) unificar, no mesmo banco relacional os dados coletados pelos formulários do programa *Tem Base!*, infraestruturas de pesquisa e de unidades EMBRAPII; (ii) materializar a camada de classificação por missão diretamente no banco, tornando-a reutilizável pelo dashboard; e (iii) desenvolver uma plataforma web com painel de exploração, painel administrativo e integração com o serviço de *matchmaking* de pesquisadores. A seguir são descritas, de forma organizada, as principais frentes de contribuição.

3.1 Construção do banco relacional e integração dos dados do SIMI

Partindo do dataset do SIMI, já estruturado no semestre anterior, foi construído um banco relacional unificado, em SQLite, capaz de materializar a visão integrada das organizações e servir como base operacional para a plataforma *Tem Base!*.

O primeiro passo consistiu no desenho de um esquema relacional em que a tabela `atores` ocupa o papel central do modelo, reunindo os principais atributos escalares de cada organização (nome, cidade, UF, site, contatos, descrição institucional, categoria, estágio de maturidade, variáveis de investimento, entre outros). A partir dessa estrutura nuclear, elementos multivalorados foram normalizados em dimensões próprias. Para cada um desses catálogos, foram definidas tabelas auxiliares específicas e tabelas de relacionamento que conectam as organizações aos respectivos itens, tornando explícitas relações que, nos dados originais, apareciam apenas como listas de textos separadas por vírgulas ou ponto e vírgula.

Com essa modelagem, o conjunto de informações do SIMI passa a estar disponível em um banco que favorece análises estruturadas, simplifica a evolução do esquema e viabiliza a integração com novas fontes de dados. Assim, sobre esse banco consolidado, foi criada uma view `vw_dashboard_organizacoes`, responsável por combinar, em uma única consulta, os atributos escalares e as listas de segmentos, tecnologias, áreas de expertise e demais características necessárias para alimentar a dashboard desenvolvida.

3.2 Integração dos formulários do programa *Tem Base!*

A etapa seguinte consistiu em integrar ao banco relacional, construído a partir do SIMI, os dados coletados pelos formulários do programa *Tem Base!*, em particular, o formulário de infraestruturas de pesquisa e o formulário das unidades EMBRAPII.

Em ambos os casos, o ponto de partida foi o reaproveitamento máximo das colunas já existentes na tabela principal de organizações. Campos como nome, cidade, e-mail, telefone, descrição institucional, categoria e tipo de organização foram preenchidos a partir das respostas dos formulários, garantindo que laboratórios, INCTs e unidades EMBRAPII começassem a aparecer na base integrada com o mesmo nível de detalhe das organizações originárias do SIMI. Para preservar a rastreabilidade, cada registro recebeu também uma indicação de origem, permitindo filtrar e comparar facilmente as diferentes fontes no dashboard.

As perguntas mais específicas dos formulários foram tratadas de forma a não se perder a informação das respostas. Assim, sempre que possível, essas informações foram mapeadas para dimensões já presentes na base; quando isso não era viável, foram criadas novas colunas no banco para armazenar atributos adicionais que não se encaixavam nas colunas existentes, mantendo a estrutura compatível com o restante do modelo e, ao mesmo tempo, evitando simplificações excessivas presentes nos dados dos formulários (por exemplo, condensar várias respostas em campos genéricos de texto livre).

Do ponto de vista operacional, essa integração foi implementada por meio de scripts que leem os arquivos dos formulários, transformam as respostas em valores padronizados (limpando duplicidades e variações de grafia) e atualizam o banco relacional. Esses scripts foram construídos com modos de teste e de aplicação efetiva, permitindo validar previamente o impacto da importação e evitar duplicação de registros quando uma infraestrutura ou unidade já está parcialmente cadastrada. Com isso, os formulários do programa *Tem Base!* passaram a alimentar diretamente a base integrada de organizações, enriquecendo o retrato do ecossistema com informações detalhadas sobre infraestruturas de pesquisa e unidades EMBRAPII sem quebrar a coerência do esquema existente.

3.3 Enriquecimento semântico

Para operacionalizar a perspectiva da hélice tríplice na base de dados, foram introduzidos dois atributos centrais: `setor` (setor institucional) e `tipo_organizacao` (papel funcional no ecossistema). Esses atributos foram construídos a partir das informações já presentes nos dados originais, em especial das colunas de categoria da organização e de descrições institucionais.

O primeiro passo foi definir um mapeamento determinístico entre as categorias originais (atributo *categoria*) e um conjunto enxuto de tipos de organização. A partir desse mapeamento, as organizações foram agrupadas em três grandes tipos, alinhados à fundamentação teórica: *Empresas* (organizações privadas com lógica de mercado), *Ambientes*

de inovação (parques, incubadoras, aceleradoras, hubs e estruturas híbridas) e *Infraestruturas de pesquisa* (ICTs, laboratórios, INCTs e unidades EMBRAPII). Em paralelo, o atributo `setor` foi inferido de forma automatizada com apoio de modelos de linguagem de grande porte (LLMs). Foi desenvolvido um script que consome a API da OpenAI utilizando o modelo `gpt-4o-mini` com *structured output*. Para cada organização na base, o script realiza uma chamada individual ao modelo, fornecendo como contexto o nome, a categoria original e a descrição institucional, e solicita a classificação em um dos três valores possíveis para `setor` (primeiro, segundo ou terceiro setor). O uso de saída estruturada garante que a resposta do modelo siga um esquema pré-definido, permitindo incorporar diretamente a classificação de `setor` ao conjunto de dados.

3.4 Classificação assistida por modelos de linguagem

Embora a base integrada reúna dados provenientes de diferentes fontes, todas elas têm em comum o fato de terem sido coletadas por meio de formulários, em grande parte com campos de texto livre. Isso resultou em elevada variabilidade de preenchimento, com diferenças de terminologia, granularidade e qualidade das respostas entre organizações e entre fontes (SIMI, laboratórios, INCTs, unidades EMBRAPII). Diante dessa heterogeneidade, optou-se por empregar modelos de linguagem de grande porte para apoiar a padronização e a classificação das organizações segundo as missões da Nova Indústria Brasil (NIB), em complemento às regras determinísticas definidas para o atributo `tipo_organizacao`.

Além do atributo `setor`, já mencionado anteriormente e inferido com apoio de LLMs, foi desenvolvido um *pipeline* específico para a atribuição de missões NIB às organizações. Esse *pipeline* utiliza o modelo `gpt-4o-mini`, por meio da API da OpenAI, com *structured output*, e segue, em linhas gerais, as seguintes etapas:

1. Definição de um *prompt* estruturado que apresenta, para cada organização, um resumo rico (*nome*, categoria, segmentos, tecnologias, áreas de expertise, programas e descrição) e descreve em linguagem natural as seis missões da NIB, incluindo "como a organização pode se encaixar" em cada missão;

2. Instrução explícita para que o modelo retorne um objeto JSON estruturado contendo o `org_id`, a lista de missões selecionadas, `confidence` (entre 0 e 1) e uma justificativa textual para cada missão atribuída;

3. Validação programática das respostas (checagem de sintaxe do JSON, consistência do `org_id`, restrição do conjunto de missões aos rótulos válidos, limites para o número de missões por organização) e registro das classificações em uma tabela dedicada (`organizacao_missoes_nib`), com armazenamento da resposta bruta do modelo para fins de auditoria e reprocessamento futuro.

O uso de *structured output* com o modelo `gpt-4o-mini` permite integrar diretamente a classificação automatizada ao banco relacional, preservando o vínculo entre organização, missões atribuídas, grau de confiança e justificativas em linguagem natural. Além disso, *pipelines* análogos foram empregados para normalizar e classificar os valores das colunas `segmento_atuacao` e `tecnologias`, reduzindo a variabilidade dos rótulos provenientes dos formulários originais. Esse *pipeline* de LLM tem, assim, dupla função: por um lado, fornece uma classificação de referência, auditável e explicada; por outro, serve como apoio à curadoria humana na revisão de casos ambíguos ou situados na fronteira entre diferentes

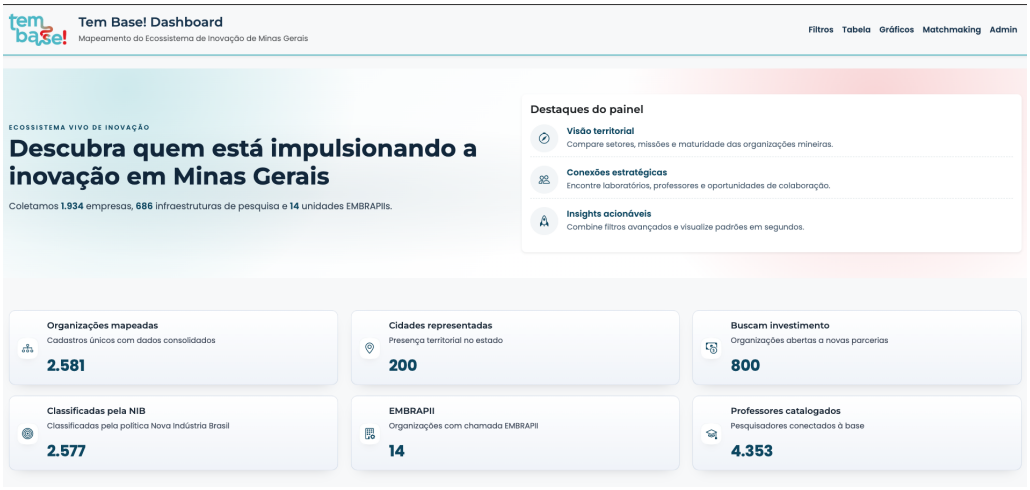

Figura 1: Tela inicial da dashboard

missões, bem como no refinamento incremental das regras de classificação ao longo do tempo.

3.5 Desenvolvimento da plataforma *Tem Base!*

A terceira frente de trabalho corresponde ao desenvolvimento da plataforma *Tem Base!*, um painel interativo em Dash (sobre Flask, em Python) para explorar o ecossistema de inovação mineiro, alimentado por uma API que expõe o banco relacional. Sua arquitetura é modular: um módulo de dados (`data_utils`) gerencia o consumo dessa API, a verificação de tabelas e *views* e a configuração dinâmica de colunas que ajusta o layout conforme os campos disponíveis; a camada de interface (`layouts`) organiza o front-end em seções de abertura, filtros, métricas, tabela detalhada, gráficos temáticos, área de professores, *matchmaking* e rodapé; e a lógica reativa (`callbacks`) conecta filtros, métricas, gráficos e tabela, atualizando estatísticas e a visão detalhada das organizações em função das escolhas do usuário. A aplicação também inclui autenticação e proteção de rotas sensíveis com sessão autenticada.

3.6 Módulo de *matchmaking* de pesquisadores

O quarto eixo de contribuição é o desenvolvimento de um módulo dedicado a aproximar pesquisadores com perfis similares, combinando embeddings textuais (títulos e resumos de artigos) com embeddings de grafo de coautoria. Esse módulo é exposto como um serviço independente de *matchmaking*, acessível via API pelo dashboard.

Dessa forma, o mapeamento institucional do ecossistema é conectado à dimensão de competências individuais de docentes e pesquisadores, sustentando serviços como o *MatchDay* e a identificação de parceiros acadêmicos em temas específicos.

4 Conclusões e Trabalhos Futuros

Este trabalho resultou no mapeamento de 2 581 organizações do ecossistema de inovação em Minas Gerais, consolidadas em um banco relacional a partir de diferentes fontes: 2 409 registros provenientes da base SIMI, 158 de formulários de infraestruturas de pesquisa e

Figura 2: Interface do matchmaking

14 de formulários Embrapii. Sobre esse universo, 2 577 organizações receberam ao menos uma missão da Nova Indústria Brasil, com uma distribuição final de 1 934 organizações do Segundo Setor (empresas privadas), 382 do Primeiro Setor (governo e universidades) e 265 do Terceiro Setor (organizações da sociedade civil), além de 1 728 empresas, 686 infraestruturas de pesquisa e 167 ambientes de inovação. A partir dessa base estruturada foram desenvolvidos o dashboard modular (*Tem Base!*) e o serviço de *match-making* entre pesquisadores, combinando práticas de engenharia de dados, modelos de linguagem e visualização interativa. Do ponto de vista técnico, as principais contribuições incluem: (i) a definição de um modelo relacional extensível para organizações de PD&I e suas classificações; (ii) a consolidação de múltiplas fontes de dados em um único banco normalizado; (iii) a aplicação prática de LLMs em um cenário real de classificação temática; e (v) a entrega de um dashboard autenticado, preparado para uso pelo BH-TEC.

Como trabalhos futuros, destacam-se algumas linhas de continuidade:

- Ampliar a cobertura de dados, incorporando novas fontes, como por exemplo, editais de fomento;
- Integrar um assistente conversacional baseado em LLM, utilizando técnicas de Retrieval-augmented generation (RAG), para possibilitar consultas em linguagem natural sobre o conjunto de dados do ecossistema mapeado.

Do ponto de vista acadêmico, o projeto ilustra o potencial de combinar práticas modernas de engenharia de dados com modelos de linguagem e técnicas de recomendação para apoiar políticas públicas de inovação.

Referências

[1] BH-TEC. "BH-TEC lança programa inédito para conectar pesquisa e setor produtivo e inicia série de webinares exclusivos," 2025. Disponível em: https://bhtec.org.br/2025/11/04/bh-tec-lanca-programa-inedito-para-conectar-pesquisa-e-setor-produtivo-e-inicia-serie-de-webinares-exclusivos/. Acesso em: 29 nov. 2025.

[2] ETZKOWITZ, H.; LEYDESDORFF, L. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, v. 29, n. 2–3, p. 109–123, 2000.

[3] BRASIL. Conselho Nacional de Desenvolvimento Industrial. *Nova Indústria Brasil: política industrial para a neoindustrialização*. Brasília, 2024.