

# EsqueMapa: Exploração Temática e Topológica de Esquemas de Bancos de Dados Relacionais

Allan Miller Silva Lima<sup>1</sup>, Rui Nóbrega de Pontes Filho<sup>1</sup>, Samuel Henrique Santos Silva<sup>1</sup>,  
Denis Mayr Lima Martins<sup>2</sup> and Fernando Buarque de Lima Neto<sup>1</sup>

<sup>1</sup>Departamento de Engenharia da Computação – Universidade de Pernambuco  
Recife – PE – Brasil

<sup>2</sup>Departamento de Computação e Matemática – Universidade de São Paulo  
Ribeirão Preto – SP – Brasil.

{amsl, rnpf, shss, fb1n}@ecom.poli.br, martins.denis@usp.br

**Abstract.** *O entendimento e a navegação de esquemas de bancos de dados complexos oferecem desafios críticos de interpretabilidade e descoberta de padrões. Apresentamos o EsqueMapa<sup>1</sup>, um sistema que combina modelos de linguagem (LLMs) para extração contextualizada de temas e Mapas Auto-Organizáveis (SOM) para mapear a estrutura latente do esquema. Para mitigar alucinações, implementamos validação baseada em métricas de consistência léxica e semântica. Avaliado no repositório E-SUS APS, o sistema reconstrói automaticamente a arquitetura temática oficial e identifica padrões emergentes não documentados. Resultados obtidos demonstram que EsqueMapa é uma ferramenta interativa e eficaz para navegação e exploração esquemas complexos.*

**Abstract.** *Understanding and navigating complex database schemas pose critical challenges for interpretability and pattern discovery. We present EsqueMapa<sup>1</sup>, a system that integrates Large Language Models (LLMs) for contextualized table-topic extraction with Self-Organizing Maps (SOM) to map latent schema structures. To mitigate LLM hallucinations, we employ validation via lexical and semantic consistency metrics. Evaluated on the E-SUS APS repository, EsqueMapa automatically reconstructs official thematic structures and identifies undocumented patterns. Our results demonstrate that EsqueMapa is an effective tool for iteratively navigating and exploring complex schemas.*

## 1. Introdução

A democratização do acesso a grandes repositórios de dados públicos e organizacionais impulsiona uma demanda crescente por ferramentas que permitam a interação eficaz entre usuários não-técnicos e estruturas de dados complexas. Contudo, a escalabilidade dos esquemas modernos, frequentemente compostos por centenas ou milhares de tabelas acumuladas, cria um gargalo na compreensão semântica e temática para analistas de diversas áreas [Glenis and Koutrika 2021]. O mapeamento manual do conteúdo de cada entidade torna-se inviável em escala, exigindo abordagens automatizadas capazes de inferir significado semântico diretamente de metadados técnicos.

<sup>1</sup>Vídeo de demonstração: <https://youtu.be/18CEIpjntRc>

Neste cenário, *Large Language Models* (LLMs) apresentam-se como uma fonte promissora de conhecimento adicional para inferir o significado inerente ao esquema de um banco de dados, especialmente em domínios públicos e abertos. No entanto, a aplicação direta de LLMs carece de avaliação sistemática sobre sua confiabilidade, frente a problemas de alucinação ou interpretações equivocadas de esquemas complexos.

Para mitigar essas limitações, este trabalho introduz o *EsqueMapa*, um sistema automatizado de *data profiling* que integra LLMs para extração contextualizada com técnicas de visualização não supervisionada. O núcleo do *EsqueMapa* é um *pipeline* híbrido que combina a capacidade linguística dos LLMs para extração semântica de temas ligados ao esquema do banco de dados com *Self-Organizing Map* (SOM) para agrupamento e navegação interativa dos temas resultantes. Diferentemente de abordagens existentes, o sistema implementa validações rigorosas de confiabilidade através da comparação inter-modelo e visualização macroestrutural de temas, garantindo a interpretabilidade das descobertas.

A implementação é demonstrada aplicada ao repositório do E-SUS APS, onde validamos a capacidade do sistema em revelar padrões temáticos ocultos em grandes volumes de dados. Os resultados evidenciam que a exploração de metadados pode ser tornada interativa e interpretável para analistas, sem sacrificar a precisão técnica. O código fonte está disponível em <https://github.com/allanmiller143/esquemapa>.

## 2. Trabalhos Relacionados

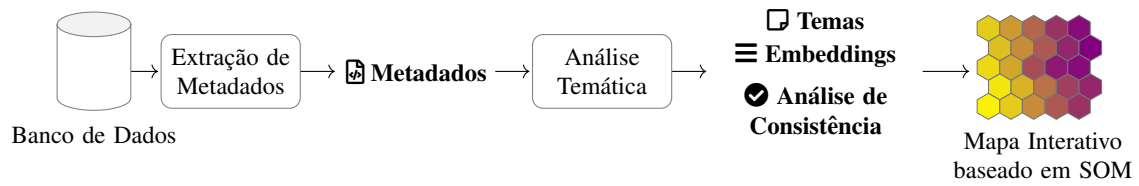
Trabalhos recentes em *data profiling* exploram o uso de LLMs para enriquecimento de metadados (e.g., LLMDap [Jiang et al. 2025], que promove anotação automática de dados científicos) e para extração e análise de documentos contratuais [Han and Jalagam 2025]. Estes trabalhos demonstram o potencial de LLMs para compreensão contextual, mas carecem de validação sistemática de confiabilidade em contextos variados.

Estudos recentes têm se dedicado a reduzir o esforço necessário para navegar dados do SUS via técnicas de visualização de dados e de AI. Este é o caso do sistema *SmartApSUS* [Pinheiro et al. 2025] cujo objetivo é auxiliar na tomada de decisão no contexto da atenção primária à saúde. O sistema utiliza mecanismos de otimização para realizar a alocação de unidades e de profissionais de saúde. Outra ferramenta é o *Painel de Vigilância da Saúde Materna* [Domingues et al. 2024] que permite analisar a série histórica de 30 indicadores socioeconômicos e de saúde materna em diferentes níveis de granularidade, seja por município, unidade da federação, macrorregião e país.

*EsqueMapa* distingue-se pela integração híbrida entre LLMs e técnicas de visualização, com foco específico na compreensão semântica e temática do esquema do banco de dados. Nossa proposta introduz um mecanismo de validação de confiabilidade sistemática, mitigando riscos de alucinação.

## 3. Visão Geral

O sistema proposto está organizado em um *pipeline* conforme a Figura 1. O esquema e os metadados do banco de dados de entrada são extraídos proceduralmente e analisados via LLMs. Esta análise resulta na classificação temática das tabelas inclusas no esquema, além de avaliação de consistência dos temas produzidos pelos LLMs. Os temas identificados são então agrupados via SOM para facilitar a navegação e exploração pelo usuário.



**Figura 1. Estrutura geral de processamento de metadados no *EsqueMapa*.**

**Extração e Preparação de Metadados.** Cada uma das 1.100 tabelas presentes no E-SUS APS é processada de forma a serem coletados os metadados a seguir:

- **Esquema estrutural:** nome do schema, nome da tabela, número de linhas, chave primária e chaves estrangeiras (colunas relacionadas e tabelas referenciadas).
- **Características dos atributos:** nome, tipo de dado, nulabilidade e estatísticas descritivas (e.g., valores mínimo e máximo, média e moda).
- **Exemplos:** até três linhas de exemplo para enriquecimento contextual.

Metadados são estruturados em formato JSON e podem ser vistos publicamente em [Lima et al. 2026]. Após este préprocessamento, tabelas completamente vazias são removidas, resultando em um total de 885 tabelas válidas para análise subsequente.

**Classificação Temática por LLMs.** Três modelos de LLM distintos foram utilizados: *GPT-4o-mini*, *Mistral Small* e *Deepseek-chat*. Para garantir a estabilidade das respostas, cada modelo é executado três vezes de forma independente sobre o mesmo conjunto de dados, totalizando nove execuções completas. O prompt solicitou a sugestão de até três temas de relevância por tabela e está disponível publicamente em [Lima et al. 2026]. Todos os modelos são configurados com temperatura zero para maximizar o determinismo e reduzir a variabilidade nas respostas [Renze 2024]. As respostas produzidas são formatadas em JSON contendo os temas identificados.

**Análise de consistência.** A confiabilidade das classificações foi medida por métricas de consistência léxica e semântica em níveis intra e inter-modelo. A consistência léxica utilizou o coeficiente de Sørensen-Dice para ajuste exato, enquanto a semântica empregou uma variante *soft* com *embeddings* via *OpenAI text-embedding-3-large*. Estes vetores foram armazenados localmente (*.npy*) para reuso na etapa seguinte, esses temas foram pareados via algoritmo húngaro para capturar similaridades conceituais. O protocolo de três execuções ( $n = 3$ ) permitiu validar a robustez das inferências e detectar a variabilidade dos modelos com baixo custo computacional.

**Agrupamento via Self-Organizing Map.** Após a classificação temática e análise de consistência, os temas identificados são agrupados por um SOM para revelar padrões macro-estruturais da base de dados. SOM é uma rede neural não supervisionada projetada para redução de dimensionalidade e visualização de dados complexos. O algoritmo mapeia um espaço de alta dimensão em uma grade bidimensional, preservando a topologia dos dados originais durante o processo de treinamento competitivo. No contexto do *EsqueMapa*, o SOM é aplicado sobre as representações vetoriais extraídas pelos LLMs para agrupar automaticamente tabelas com semântica similar, facilitando a navegação interativa e a identificação de clusters temáticos no esquema.

Para treinar o SOM, uma busca exaustiva foi realizada para seleção de hiperparâmetros tais como: tamanho do mapa, taxa de aprendizado  $\alpha$ , raio inicial de vizinhança  $\sigma$  e número de épocas de treinamento. A qualidade de cada configuração de hiperparâmetros foi avaliada a partir de uma função objetivo composta, que considera simultaneamente o erro de quantização e o erro topográfico para avaliar, respectivamente, a representação dos dados e a preservação da estrutura topológica [Vesanto and Alhoniemi 2000]. A melhor configuração encontrada consistiu em um mapa  $18 \times 16$  em topologia hexagonal,  $\sigma = 2.5$ ,  $\alpha = 0.1$  e 2000 épocas. Após a convergência, cada tema foi atribuído ao seu BMU, seguindo estratégias de rotulação pós-treinamento tradicionais da literatura [Khacef et al. 2019]. Em seguida, a *U-matrix* [Vesanto and Alhoniemi 2000] é computada para cada neurônio como a distância euclidiana média de seus seis vizinhos na topologia hexagonal, revelando fronteiras entre clusters.

**Interface com usuário.** O último passo do *pipeline* transforma os resultados em uma representação gráfica intuitiva para identificação de *clusters* temáticos e análise de similaridade entre metadados. Através de um mapa de baixa dimensão que preserva relações topológicas, o sistema consolida a organização semântica do banco de dados. Esta estrutura visual é essencial para a descoberta de macrotemas emergentes e para a estruturação do conhecimento extraído do esquema original, facilitando a exploração de padrões complexos em uma interface unificada.

A interface central consiste em um mapa interativo composto por uma grade hexagonal colorida no mesmo formato da *U-matrix*, onde cada célula representa um neurônio do SOM. Os usuários podem interagir diretamente com a visualização; ao posicionar o cursor sobre um hexágono, um *tooltip* dinâmico detalha os temas agrupados e suas respectivas frequências. Essa funcionalidade permite uma inspeção granular e imediata dos dados, revelando a composição interna de cada agrupamento sem a necessidade de navegação entre diferentes telas, o que otimiza o fluxo de análise exploratória.

Para aprofundar a investigação, a interface incorpora ferramentas de filtragem por macrotemas e um campo de busca por termos-chave, que destacam automaticamente os neurônios relevantes no mapa. Complementarmente, uma barra lateral dedicada exibe a lista completa de temas do neurônio selecionado, proporcionando uma visão textual consolidada que auxilia na navegação. Tais recursos elevam a visualização espacial a uma ferramenta de consulta, permitindo que o usuário possa compreender a essência temática do banco de dados de forma intuitiva, mesmo sem qualquer contato prévio com sua complexa estrutura técnica subjacente, o mapa interativo pode ser visto na figura 2.

## 4. Resultados

Os resultados evidenciados pela tabela 1, demonstram alta estabilidade intra-modelo, com coeficientes de *Soft Dice* superiores a 0,91, destacando o modelo GPT-4o-mini (0,951) como o mais consistente. A diferença média de +0,14 pontos entre as métricas léxica e semântica confirma que os modelos frequentemente utilizam variações de termos para conceitos equivalentes. Já a consistência inter-modelo apresentou valores inferiores (*Soft Dice* médio de 0,71), indicando que, embora os modelos concordem semanticamente em cerca de 70% dos temas, utilizam vocabulários distintos para expressá-los. O par OpenAI-Mistral obteve o maior alinhamento conceitual (0,731), reforçando que a equivalência

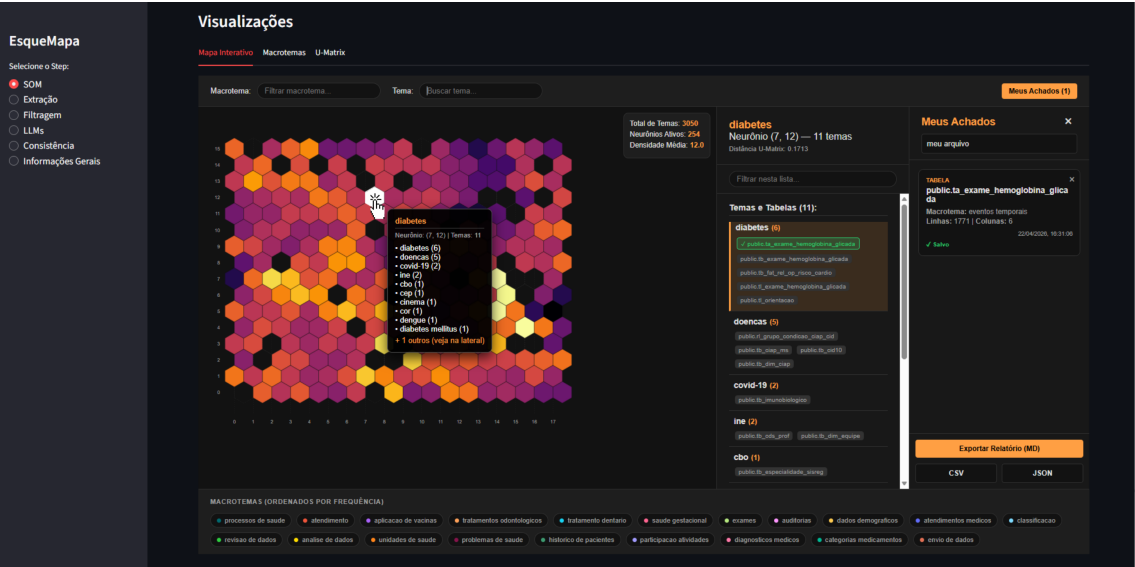


Figura 2. Visão geral da interface com o usuário incluindo o SOM interativo.

temática não implica identidade léxica e validando a importância de métricas semânticas na avaliação de classificações textuais.

Tabela 1. Consistência intra e inter-modelo (885 tabelas).

| Modelo / Par       | Dice Léx. | Dice Sem.    |
|--------------------|-----------|--------------|
| GPT-4o-mini        | 0.862     | <b>0.951</b> |
| DeepSeek Chat      | 0.797     | 0.934        |
| Mistral Small      | 0.766     | 0.916        |
| OpenAI ↔ Mistral   | 0.286     | <b>0.731</b> |
| OpenAI ↔ DeepSeek  | 0.219     | 0.703        |
| Mistral ↔ DeepSeek | 0.216     | 0.695        |

A aplicação do SOM sobre os temas extraídos resultou na identificação de 254 temas granulares, que foram organizados em clusters espaciais refletindo a estrutura do E-SUS APS. O critério de sucesso foi a capacidade do *framework* em reconstruir autonomamente a arquitetura de dados oficial, composta por 14 macrotemas: (1) Cadastro Individual, (2) Cadastro Domiciliar, (3) Atendimento Individual, (4) Atendimento Odontológico, (5) Atividade Coletiva, (6) Procedimentos, (7) Visita Domiciliar, (8) Marcadores de Consumo Alimentar, (9) Avaliação de Elegibilidade, (10) Atendimento Domiciliar, (11) Síndrome Neurológica por Zika/Microcefalia, (12) Vacinação, (13) Cuidado Compartilhado e (14) IVCF-20.

A análise dos agrupamentos revelou uma alta densidade de neurônios em áreas críticas como *Vacinação* e *Atendimento Individual*, demonstrando que o SOM preservou corretamente as relações de proximidade semântica entre tabelas correlatas. Além da convergência com os domínios oficiais, o mapa permitiu a identificação de temas emergentes não explicitamente documentados como macrotemas, mas presentes no esquema, como *Saúde Gestacional* e *Gestão de Insumos*. A visualização via U-Matrix confirmou a formação de fronteiras claras entre clusters distintos, validando a eficácia do sistema em

reduzir a complexidade do esquema original para uma estrutura navegável e semanticamente coerente para o analista.

## 5. Conclusão

Este trabalho apresenta o *EsqueMapa*, um sistema automático de análise temática de esquemas de bancos relacionais que integra LLMs e SOMs. Validado na base de dados do E-SUS APS, o sistema proposto identifica padrões ocultos e confirma convergência com domínios oficiais (e.g., vacinação, odontologia, procedimentos) além de revelar temas emergentes não documentados, como saúde gestacional e exames. Trabalhos futuros incluem substituir *embeddings* genéricos por modelos de domínio médico específico e validar a generalização dos resultados para outras áreas e contextos institucionais, além de iniciar testes de usabilidade com usuarios reais.

## Agradecimentos

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – código 001. Os autores também reconhecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE).

## Referências

- Domingues, R. M. S. M. et al. (2024). Painel de vigilância da saúde materna: ferramenta para vigilância epidemiológica. *Rev. Bras. Epidemiol.*, 27:e240009.
- Glenis, A. and Koutrika, G. (2021). Pyexplore: Query recommendations for data exploration without query logs. In *Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21*, page 2731–2735, New York, NY, USA. Association for Computing Machinery.
- Han, C. and Jalagam, S. (2025). Metadata extraction leveraging large language models. arXiv preprint arXiv:2510.19334.
- Jiang, S., Sørbø, S., Tinn, P., Karim, S. F., and Roman, D. (2025). Llmmap: Llm-based data profiling and sharing. In *VLDB 2025 Workshop: 3rd Data Economy Workshop (DEC)*.
- Khacef, L., Miramond, B., Barrientos, D., and Upegui, A. (2019). Self-organizing neurons: toward brain-inspired unsupervised learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.
- Lima, A., Filho, R., Silva, S., Martins, D., and Neto, F. (2026). Prompt de entrada para esquemapla. <https://doi.org/10.5281/zenodo.18299520>.
- Pinheiro, T. et al. (2025). Smartapsus: Integrando dados, previsao e otimizacao para fortalecer a gestao da aps no brasil. In *Proc. SBBD*, pages 82–87. SBC.
- Renze, M. (2024). The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 7346–7356. Association for Computational Linguistics.
- Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. In *IEEE Trans. Neural Networks*, volume 11, pages 586–600.