

AedesTAG: vigilância de arboviroses com documentos semanticamente estruturados e *Retrieval-Augmented Generation*

Renato B. D. de Oliveira^{1,2}, Solange Nice Alves-Souza²

¹Instituto Federal de São Paulo – Guarulhos, Brasil

²Departamento de Engenharia da Computação
Universidade de São Paulo (USP) – São Paulo, SP – Brasil

{renato.bueno, ssouza}@usp.br, renato.bueno@ifsp.edu.br

Abstract. *This article presents a Retrieval-Augmented Generation (RAG) approach for question answering in arbovirus surveillance. The proposed workflow transforms structured records from the Brazilian Notifiable Diseases Information System (SINAN) into semantically organized textual documents enriched with metadata, making epidemiological evidence retrievable in a contextualized and explainable way within RAG systems. Preliminary experiments with dengue data show that semantically structured documents facilitate the contextual retrieval of epidemiological evidence, supporting more interpretable answers.*

Resumo. *Este artigo apresenta uma abordagem de Retrieval-Augmented Generation (RAG), ou seja, geração aumentada por recuperação de informação para perguntas e respostas em vigilância de arboviroses. O fluxo de trabalho proposto transforma registros estruturados do sistema de informação de agravos de notificação (SINAN) em documentos textuais semanticamente organizados e enriquecidos por metadados, tornando evidências epidemiológicas recuperáveis de forma contextualizada e explicável para o RAG. Experimentos preliminares com dados de dengue mostram que documentos semanticamente estruturados facilitam a recuperação contextual de evidências epidemiológicas, favorecendo respostas mais interpretáveis.*

1. Introdução

No Sistema de Informação de Agravos de Notificação (SINAN), milhares de registros epidemiológicos são organizados em formato tabular para fins de notificação e monitoramento. Entretanto, essa estrutura dificulta sua utilização direta em mecanismos de recuperação semântica voltados a perguntas em linguagem natural, pois dados tabulares exigem interpretação adicional de linhas, colunas, agregações e relações entre atributos antes de se tornarem unidades semanticamente recuperáveis [de Souza et al. 2024].

Os registros brutos de notificação do SINAN frequentemente carecem de organização contextual voltada à recuperação de informação para resposta a perguntas, especialmente quando as questões analíticas envolvem padrões de sintomas, sorotipos da dengue, evolução temporal da doença, comparações espaciais ou interpretação da qualidade dos dados e da sua completude.

Embora abordagens como *Text-to-SQL* permitam converter perguntas em consultas estruturadas, sua eficácia depende da identificação explícita de elementos do esquema relacional, como tabelas, colunas, filtros e agregações [Liu et al. 2026, de Souza et al. 2024]. No domínio epidemiológico, perguntas interpretativas frequentemente exigem a combinação de múltiplas evidências distribuídas em diferentes atributos e agregações, além de contexto semântico adicional, o que limita respostas baseadas apenas em consultas estruturadas.

Diante disso, este trabalho propõe uma abordagem de *Table-Augmented Generation* (TAG) baseada em documentos semanticamente estruturados para apoio à resposta a perguntas em vigilância de arboviroses. Em vez de indexar diretamente os registros brutos do SINAN, a proposta transforma notificações epidemiológicas em documentos textuais semanticamente organizados e enriquecidos com metadados, representando diferentes perspectivas analíticas, como total de casos, sintomas, sorotipos e sinais temporais. O objetivo é tornar a evidência epidemiológica mais acessível aos mecanismos de recuperação e aos modelos de linguagem, favorecendo respostas mais contextualizadas, interpretáveis e rastreáveis.

Como contribuição, o trabalho apresenta uma estratégia de transformação de dados tabulares em documentos semanticamente estruturados e um fluxo de recuperação que distingue perguntas determinísticas e não determinísticas no domínio da dengue.

2. Trabalhos relacionados

Abordagens recentes têm explorado o uso de LLMs para atribuir semântica a dados estruturados em tabelas. O estudo de [Conrado et al. 2025], por exemplo, utiliza modelos de linguagem para identificar relações semânticas entre atributos de registros clínicos eletrônicos, permitindo a construção de grafos de conhecimento. Esse tipo de abordagem evidencia que uma camada semântica adicional pode tornar dados tabulares mais interpretáveis em tarefas analíticas e de recuperação de informação.

No contexto do RAG, representações textuais derivadas de dados estruturados também têm sido utilizadas para ampliar a capacidade de recuperação contextual. Trabalhos recentes mostram que a transformação de tabelas em representações semanticamente interpretáveis facilita a indexação vetorial e a recuperação de conteúdo relevante em perguntas e respostas baseadas em LLMs [de Souza et al. 2024]. Além do conteúdo textual, o enriquecimento com metadados contribui para melhorar a rastreabilidade das respostas e reduzir ambiguidades entre documentos semanticamente semelhantes [Mukherjee 2025, Brown et al. 2025, Chen and Taipalus 2025].

Entretanto, mesmo em pipelines RAG, limitações permanecem quando informações relevantes são fragmentadas durante a etapa de *chunking*, o que pode comprometer a recuperação contextual e favorecer respostas inconsistentes ou incompletas [Mukherjee 2025]. Em vigilância epidemiológica, perguntas como qual sorotipo predomina em determinada região? podem ser tratadas por filtros estruturados sobre documentos agregados, enquanto perguntas como há aumento de casos graves? exigem combinar múltiplas evidências de tempo, local e características epidemiológicas.

Diferentemente dessas abordagens, este trabalho propõe a geração de documentos semanticamente estruturados a partir de registros epidemiológicos do SINAN, combinando texto em linguagem natural e metadados epidemiológicos organizados para

recuperação contextualizada em RAG. O objetivo não é apenas permitir a recuperação dos dados, mas estruturar evidências epidemiológicas de forma adequada tanto para perguntas determinísticas, como “quantos casos foram registrados por UF?”, quanto para perguntas interpretativas, como “há aumento de casos graves em determinada região?”

3. Arquitetura da solução

A Figura 1 apresenta o fluxo de dados da proposta, organizado em três fases principais: A – ingestão de dados, B – processamento da pergunta e C – geração da resposta. O fluxo foi concebido para transformar registros epidemiológicos originalmente tabulares em documentos semanticamente estruturados, permitindo seu uso em mecanismos de recuperação contextual e geração de respostas.

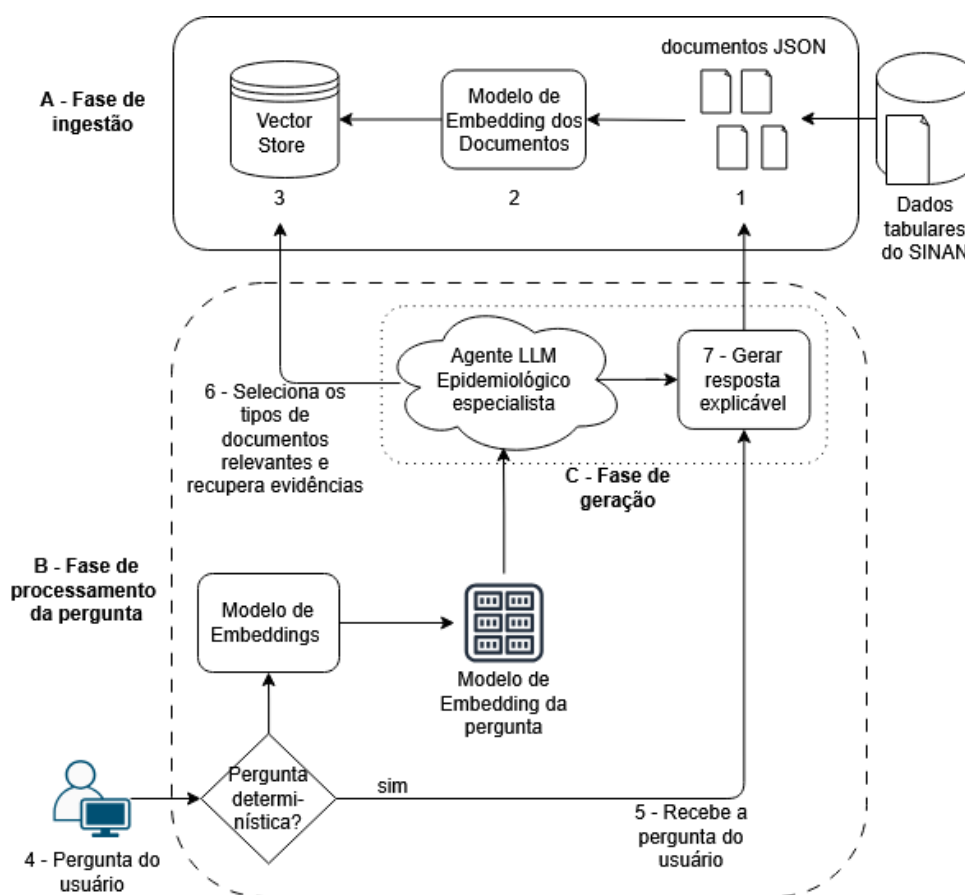


Figura 1. Visão detalhada do fluxo de geração de resposta ao usuário

3.1. A - Ingestão de dados

Os dados utilizados são provenientes do SINAN, originalmente disponibilizados em formato tabular. Os arquivos podem ser obtidos por meio do portal do Ministério da Saúde ou acessados com apoio da biblioteca PySUS¹.

Embora a base original contenha 121 colunas, foi selecionado apenas um subconjunto de atributos epidemiologicamente relevantes para compor documentos semanticamente estruturados. Essa seleção varia conforme o tipo de informação a ser representada.

¹ <https://pysus.readthedocs.io/pt/latest/databases/SINAN.html>

No caso do documento referente ao total de casos por Unidade Federativa (UF) e ano, por exemplo, foram utilizadas três colunas da base original. Inicialmente foram criados 10 documentos, mas pretende-se expandir esse número.

O código-fonte e os arquivos utilizados encontram-se disponíveis em repositório público no GitHub². Um exemplo de documento com os sorotipos da dengue pode ser visualizado na figura 2.

```
{
  "metadata": {
    "document_type": "sorotipos_por_uf_com_codigos_e_nomes",
    "nu_ano": 2026,
    "sg_uf_not": "28",
    "uf_name": "Sergipe",
    "sorotipo": "DEN 3",
    "total_casos": 2,
    "semana_epidemiologica_range": "1 a 11"
  },
  "text_content": "No ano de 2026, no estado de Sergipe (codigo 28), foram notificados 2 casos do sorotipo DEN 3."
}
```

Figura 2. Trecho do JSON estruturado semanticamente com metadados e conteúdo. Sorotipos da dengue por semana epidemiológica e UF

Esses documentos foram concebidos para responder perguntas analíticas que não são diretamente contempladas por sistemas como o InfoDengue [Fiocruz 2025] e o painel de monitoramento de arboviroses do Ministério da Saúde [Ministério da Saúde 2025]. Como exemplo, destaca-se a possibilidade de responder perguntas como: quais sorotipos circulam em cada estado?

3.2. B - Processamento da pergunta

As perguntas submetidas ao sistema são inicialmente classificadas em dois fluxos de resposta. No primeiro, consultas associadas diretamente a metadados ou a documentos agregados específicos são tratadas por regras previamente definidas, permitindo recuperar informações objetivas por meio de filtros estruturados.

No segundo fluxo, perguntas que exigem interpretação de múltiplas evidências acionam a recuperação vetorial sobre os documentos previamente indexados. Nesses casos, são selecionados os documentos semanticamente mais próximos da consulta, que posteriormente compõem o contexto enviado ao modelo de linguagem para geração da resposta.

Nos casos interpretativos, a recuperação inicial dos documentos utiliza representações vetoriais geradas com o modelo *text-embedding-3-small*. Em seguida, os três documentos mais relevantes passam por uma etapa de reranqueamento com o modelo *gpt-4o-mini*, que prioriza os trechos mais compatíveis com o contexto da pergunta e seus metadados epidemiológicos.

3.3. C - Geração da resposta

A geração da resposta depende do tipo de pergunta formulada pelo usuário, conforme definido na fase de processamento da pergunta. Quando a pergunta é classificada como

²<https://github.com/renatobdo/arbotag.git>

não determinística, o sistema aciona o modelo de linguagem que foi denominado como Agente epidemiológico especialista, responsável por construir respostas a partir dos documentos semanticamente estruturados recuperados. O agente é instruído a utilizar apenas as evidências presentes no contexto recuperado, evitando a introdução de informações não suportadas pelos documentos disponíveis.

Para exemplificar a execução, submeteu-se a pergunta: “*Quais sorotipos circulam em São Paulo em 2026?*” O sistema recuperou o documento correspondente e gerou a resposta: “*os sorotipos que circulam em São Paulo nesse ano são DEN 1, DEN 2, DEN 3 e DEN 4*”. Essa execução demonstra a viabilidade da proposta em fornecer respostas rastreáveis, que podem ser visualizadas de forma integrada por meio de grafos de evidências como na Figura 3.

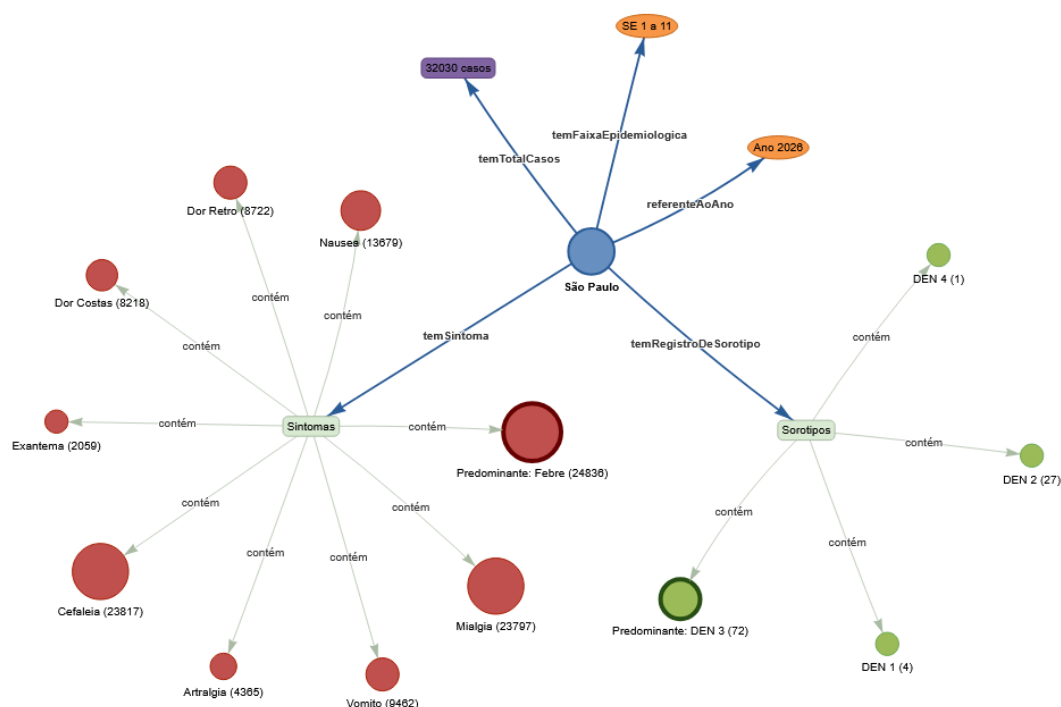


Figura 3. Grafo por UF, sorotipo e sintomas para a semana epidemiológica 1 a 11 de 2026

Contudo, observam-se limitações quando as perguntas envolvem causalidade, como “*Por que os casos aumentaram?*” Nesses casos, o sistema identifica que a informação é insuficiente e orienta o agente a explicitar a limitação, em vez de gerar inferências não fundamentadas. Isso ocorre porque o contexto atual é restrito ao SINAN, evidenciando a necessidade de integração com dados climáticos e socioeconômicos para análises mais profundas.

4. Conclusão

A utilização de documentos semanticamente estruturados enriquecidos com metadados auxilia no processo de recuperação da informação. Os resultados preliminares indicam potencial para respostas mais rastreáveis e interpretáveis. Como trabalhos futuros

pretende-se acrescentar uma ontologia integrada com virtual knowledge graphs para permitir alinhamento ontológico e maior integração com outras fontes de dados. Além disso, documentos com novos tipos de agregações ou combinações serão incorporados para responder a novas questões relacionadas à dengue.

5. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2023/10080-3. Os autores também reconhecem o uso de ferramentas de Inteligência Artificial Generativa (Gemini e ChatGPT) para o suporte na revisão gramatical do manuscrito.

Referências

- Brown, A., Roman, M., and Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *Big Data and Cognitive Computing*, 9:320. <https://doi.org/10.3390/bdcc9120320>.
- Chen, B. and Taipalus, T. (2025). Metadata-aware rag for enforcing access control and metadata-based filtering: Proof of concept and evaluation. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pages 1–7. ACM. <https://doi.org/10.1145/3769002.3769952>.
- Conrado, R. C. G., Campos, E. M., Jr., C. T., Traina, A. J. M., and Cazzolato, M. T. (2025). From tables to graphs with clinicoatlas: Leveraging llms to support modeling and mining knowledge graphs. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados (SBBd 2025)*, pages 64–69. Sociedade Brasileira de Computação - SBC. https://doi.org/10.5753/sbbd_estendido.2025.247610.
- de Souza, E. A., da Silva, P. F., Gomes, D., Batista, V., Batista, E., and Pacheco, M. (2024). Tablerag: A novel approach for augmenting llms with information from retrieved tables. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*, pages 182–191. Sociedade Brasileira de Computação. <https://doi.org/10.5753/stil.2024.245371>.
- Fiocruz (2025). Infodengue. <https://info.dengue.mat.br/services/>. Acessado em 30-03-2026.
- Liu, M., Wang, X., Xu, J., Yi, W., and Wolfson, O. (2026). A systematic review of natural language interfaces for databases. *Frontiers of Computer Science*, 20:2011623. <https://doi.org/10.1007/s11704-025-50592-w>.
- Ministério da Saúde (2025). Painel de monitoramento das arboviroses. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/aedes-aegypti/monitoramento-das-arboviroses>. Acesso em: 30 mar. 2026.
- Mukherjee, S. (2025). Leveraging metadata and llm context to overcome rag challenges and limitations. In *2025 IEEE 2nd International Conference on Big Data Science and Engineering (ICBDSE)*, pages 1–8. IEEE. <https://doi.org/10.1109/ICBDSE65491.2025.11220064>.