

Visões semânticas materializadas para representação interpretável de dados tabulares epidemiológicos

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. *Epidemiological surveillance data are commonly distributed in tabular formats, preserving structured records but hindering direct interpretation by both human analysts and language-based systems. This work proposes a semantic document engineering approach based on semantic materialized views, which transform tabular data into semantically structured textual representations enriched with metadata. The approach aggregates epidemiological evidence along dimensions such as location, time, symptoms, serotypes, and hospitalization indicators, generating interpretable intermediate representations that preserve traceability to original data sources. A prototype using arbovirus surveillance data demonstrates how the proposed representation supports both deterministic queries and interpretative question answering using retrieval-augmented generation (RAG).*

Resumo. *Dados de vigilância epidemiológica são frequentemente disponibilizados em formato tabular, preservando registros estruturados, mas dificultando sua interpretação por analistas e sistemas baseados em linguagem. Este trabalho propõe uma abordagem de engenharia de documentos semânticos baseada em visões semânticas materializadas, que transformam dados tabulares em representações textuais estruturadas e enriquecidas com metadados. A abordagem agrega evidências epidemiológicas segundo dimensões como localização, tempo, sintomas, sorotipos e hospitalização, gerando representações intermediárias interpretáveis que preservam a rastreabilidade aos dados originais. Um protótipo^{1,2} com dados de arboviroses demonstra o potencial da abordagem para suportar consultas determinísticas e perguntas interpretativas com uso de RAG.*

1. Introdução

Sistemas de vigilância epidemiológica geram grandes volumes de registros estruturados que descrevem notificações de doenças, evolução temporal e distribuição territorial. No Brasil, repositórios públicos de saúde, como o SINAN (Sistema de informação de agravos de notificações), disponibilizam conjuntos de dados em formato tabular para o monitoramento de arboviroses e outras doenças infecciosas.

¹<https://github.com/renatobdo/semantic-materialized-view>

²<https://youtu.be/PK2KhJ4Uuig>

Apesar disso, a interpretação desses dados depende de familiaridade técnica com atributos e metadados. Essa característica cria uma lacuna entre a disponibilidade dos dados e sua interpretabilidade. Embora tabelas preservem registros granulares, elas não evidenciam o significado epidemiológico em formas diretamente compreensíveis por sistemas computacionais baseados em linguagem [Liu et al. 2023].

Neste trabalho, propomos uma abordagem baseada em visões semânticas materializadas para transformar dados epidemiológicos tabulares em representações interpretáveis. Em vez de tratar cada registro como uma unidade isolada, a abordagem constrói representações que agregam evidências epidemiológicas de acordo com dimensões relevantes, como unidade federativa, período temporal, sintomas, sorotipos, perfil de hospitalização, evolução dos casos e outras.

A proposta avança em relação a abordagens baseadas apenas em RAG ao introduzir o conceito de visões semânticas materializadas como uma camada intermediária de representação. Diferentemente de abordagens que transformam dados tabulares diretamente em texto para recuperação, a proposta estrutura os dados em unidades semânticas interpretáveis, organizadas segundo dimensões epidemiológicas, permitindo maior controle sobre interpretabilidade, rastreabilidade e reutilização das evidências. Este trabalho estende uma proposta anterior [de Oliveira and Alves-Souza 2026] baseada em documentos semânticos ao formalizar explicitamente o conceito de visões semânticas materializadas como uma camada de representação intermediária e detalha seu algoritmo.

2. Trabalhos relacionados

Sistemas recentes de perguntas e respostas têm explorado diferentes estratégias para acessar e interpretar dados estruturados e não estruturados como o uso de LLMs, RAG e Table-Augmented Generation (TAG) [Yu et al. 2025, de Souza et al. 2024]. Abordagens tradicionais em interfaces de linguagem natural para bancos de dados concentram-se na tradução de consultas dos usuários em linguagens de consulta estruturadas, como SQL [Biswal et al. 2024, O'Shaughnessy 2026]. Essas abordagens frequentemente apresentam dificuldades com perguntas interpretativas que exigem síntese semântica além da representação explícita dos dados.

RAG combina recuperação baseada em similaridade com geração de texto guiada pelo contexto recuperado, permitindo que modelos de linguagem produzam respostas fundamentadas em evidências externas ao modelo [Fan et al. 2024, Arslan et al. 2024]. Entretanto, quando aplicado a dados tabulares, o RAG depende de estratégias que transformam tabelas em blocos textuais [Biswal et al. 2024, Fang et al. 2024, Long et al. 2025].

Trabalhos mais recentes têm investigado sistemas híbridos de perguntas e respostas sobre dados heterogêneos, combinando fontes textuais e tabulares [Fang et al. 2024, José et al. 2025]. Essas abordagens destacam a importância de preservar representações estruturadas ao mesmo tempo em que possibilitam o raciocínio semântico. Ainda assim, em geral, operam por meio de execução direta de consultas ou por recuperação sobre representações não estruturadas, sem definir explicitamente representações semânticas intermediárias voltadas à interpretabilidade.

Diferentemente das abordagens existentes, que focam na recuperação de informação ou na geração de respostas, este trabalho propõe uma camada intermediária

de representação semântica. Em vez de operar diretamente sobre dados tabulares, a proposta introduz visões semânticas materializadas como unidades estruturadas de evidência, projetadas para suportar interpretabilidade, rastreabilidade e integração com modelos de linguagem.

A Tabela 1 apresenta uma comparação entre diferentes abordagens. Em *Natural Language to SQL* (NL2SQL), a estrutura dos dados é preservada por meio de consultas formais sobre esquemas relacionais. Entretanto, a interpretação das respostas permanece dependente da semântica implícita das tabelas, atributos e operações de junção.

Tabela 1. Comparação entre abordagens de acesso a dados tabulares

Abordagem	Representação	Interpretabilidade	Estrutura
NL2SQL	Consulta estruturada	Baixa	Alta
RAG tradicional	Texto sequencial	Média	Baixa
Proposta	Visões semânticas	Alta	Alta

Em abordagens tradicionais de RAG, a transformação de dados tabulares em texto sequencial facilita a leitura em linguagem natural, mas pode comprometer a estrutura original dos dados e introduzir ambiguidades semânticas. A proposta deste trabalho busca equilibrar esses aspectos ao combinar metadados estruturados com representações textuais interpretáveis, preservando relações semânticas relevantes e favorecendo a interpretabilidade e a rastreabilidade dos dados, contribuindo para que modelos de linguagem gerem respostas fundamentadas nas informações recuperadas mitigando alucinações.

3. Construção das Visões Semânticas Materializadas

A Figura 1 ilustra a extração das visões semânticas materializadas dos dados brutos do SINAN.

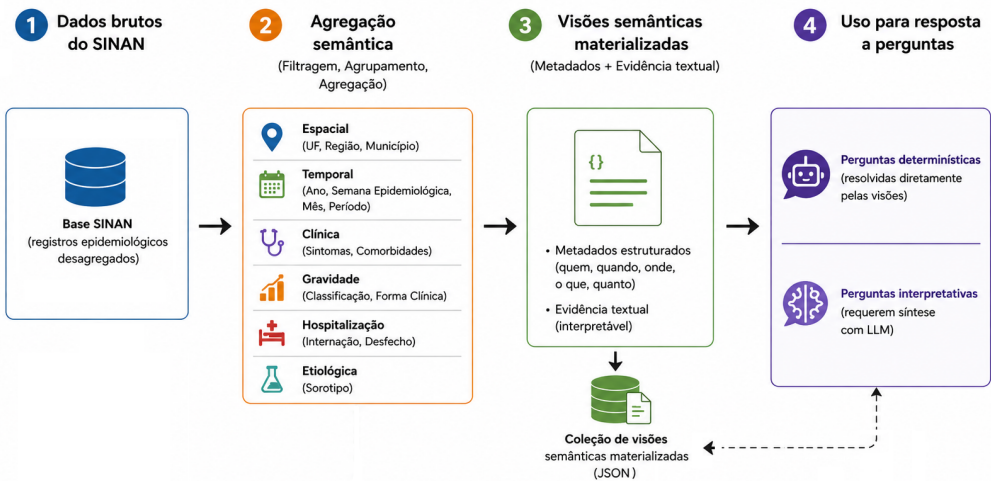


Figura 1. Processo de agregação semântica em visões semânticas materializadas para resposta a perguntas determinísticas ou interpretativas

3.1. Dados brutos do SINAN

Os dados utilizados neste trabalho são provenientes do SINAN, que armazena registros epidemiológicos estruturados em formato tabular. Cada registro corresponde a uma notificação individual, contendo atributos relacionados à identificação do caso, localização geográfica, sintomas apresentados, classificação clínica, evolução e outros indicadores relevantes.

Apesar de sua riqueza informacional, esses dados são representados por meio de códigos e variáveis categóricas que exigem conhecimento técnico para interpretação. Além disso, análises epidemiológicas frequentemente demandam múltiplas etapas de agregação, dificultando o acesso direto a informações relevantes.

3.2. Visão semântica materializada

Uma visão semântica materializada é definida como uma estrutura de representação intermediária derivada de dados tabulares, composta por dois elementos principais: (i) metadados estruturados que descrevem dimensões epidemiológicas relevantes e (ii) uma representação textual interpretável que explicita o significado dos dados agregados.

Formalmente, uma visão semântica materializada pode ser representada como:

$$V = (M, T)$$

em que M representa o conjunto de metadados estruturados e T representa o conteúdo textual derivado desses metadados. O algoritmo 1 descreve a construção de uma visão semântica materializada percorrendo etapas de agregação (*GROUP BY*), acréscimo de metadados e texto, mais detalhes estão no repositório do gihub³. Essa estrutura possibilita tanto consultas determinísticas sobre metadados quanto interpretação baseada em linguagem natural.

Algoritmo 1: Construção de uma Visão Semântica Materializada

Input: Registros epidemiológicos R , regra de agregação A , modelo de metadados M e modelo textual T

Output: Visão Semântica Materializada V

```

1 Aplicar a regra de agregação  $A$  sobre os registros  $R$ ;
2 Inicializar  $V \leftarrow \emptyset$ ;
3 foreach registro agregado  $r$  do
4   Extrair os atributos epidemiológicos relevantes;
5   Preencher o modelo de metadados  $M$ ;
6   Gerar a descrição textual utilizando o modelo  $T$ ;
7   Criar o documento semântico  $d \leftarrow \{\text{metadados}, \text{texto}\}$ ;
8   Inserir  $d$  em  $V$ ;
9 return  $V$ 
```

Diferentemente de visões materializadas tradicionais, cujo objetivo principal é otimizar consultas, as visões semânticas materializadas (VSM) introduzem uma camada de abstração voltada à interpretabilidade e ao suporte a sistemas baseados em linguagem.

³<https://github.com/renatobdo/semantic-materialized-view/tree/main/notebooks>

3.3. Demonstração

A demonstração da figura 2, evidencia como as visões semânticas materializadas estruturam a interação do usuário, em complemento ao pipeline de recuperação e geração descrito em [de Oliveira and Alves-Souza 2026]. São exibidas: (i) a lista de visões disponíveis por dimensão epidemiológica; (ii) uma pergunta determinística ou pergunta sugerida; e (iii) uma pergunta interpretativa (ex.: "Há sinais de agravamento em SP?"), cuja resposta sintetizada por LLM é exibida junto à evidência e aos metadados de origem.

Agente Inteligente de Arboviroses

Consulta semântica com filtros por: document, type, Runco determinísticos para consultas agregadas, respostas com apoio de LLM para consultas abertas, gráficos e grafos.

Perguntas sugeridas

Escolha uma pergunta sugerida

Quais sintomas predominam por estado?

Usar pergunta sugerida Consultar Limpar pergunta

Digite sua pergunta:

Há sinais de agravamento em SP? Press Enter to apply

Document types encontrados:

```

1 {
  2 "hospitalizacao_por_estado"
  3 "notificacao_por_estado"
  4 "sorotipo_por_estado"
  5 "sorotipo_por_estado_e_semana"
  6 "sintomas_por_estado"
  7 "total_casos_estado"
}

```

Pergunta atual

Há sinais de agravamento em SP?

Resposta

Com base no contexto fornecido, há indícios de um cenário preocupante em São Paulo em 2026, considerando os seguintes pontos:

- Total de Casos de Dengue:** Foram notificados 12.030 casos de dengue, com um sintoma principal de febre em 24.835 casos. Isso indica uma alta incidência da doença.
- Hospitalizações:** No período de 1 a 14 semanas epidemiológicas, foram registrados 1.280 casos hospitalizados, o que representa uma taxa bruta de internação de 3,35% e uma taxa válida de 4,18%. Essas taxas, embora não extremamente altas, indicam que uma parte significativa dos casos requer atenção médica.
- Surtos de Sorotipo:** Na semana epidemiológica 202604, foram notificados 11 casos de dengue com o sorotipo DEN 3, e na semana 202605, 13 casos do mesmo sorotipo. A presença do sorotipo DEN 3 pode ser um fator de risco, pois está associado a formas mais graves da doença.
- Taxa de Incompletude:** A taxa de incompletude dos dados foi de 19,85%, o que pode indicar que os números reais de casos e hospitalizações podem ser ainda maiores.

Esses fatores, em conjunto, sugerem um cenário de agravamento da situação de dengue em São Paulo, especialmente considerando a alta quantidade de casos notificados e a presença de um sorotipo que pode ser mais virulento.

Trecho 3

Arquivo origem: sorotipo_by_state_and_week_ag_documents.json

Document type: sorotipo_by_state_and_week

Metadados completos:

```

{
  "granularidade": "semanal",
  "ano_normalizado": "",
  "arquivo_origem": "sorotipo_by_state_and_week_ag_documents.json",
  "document_type": "sorotipo_by_state_and_week",
  "sorotipo": "DEN 3",
  "semana_normalizada": "202605",
  "uf_name": "São Paulo",
  "uf_semana_notificacao": "202605",
  "num_casos": 13,
  "categoria": "sorotipo",
  "sorotipo_normalizado": "DEN 3",
  "uf_normalizada": "São Paulo",
  "sistema_notificacao": "1",
  "justificativa_notificacao": {
    "1": "passa_documento_tipo1_0",
    "2": "passa_documento_tipo1_0"
  },
  "acore_tia": 0
}

```

Texto

Em São Paulo, na semana epidemiológica 202605, foram notificados 13 casos de dengue com o DEN 3.

Figura 2. Telas da aplicação: (a) consulta interpretativa; (b) resposta com auxílio do LLM; (c) Trecho da VSM utilizada na resposta ou $V=(M,T)$.

4. Conclusão

Este trabalho apresentou uma abordagem baseada em visões semânticas materializadas como uma camada intermediária de representação para dados epidemiológicos tabulares. A proposta permite estruturar evidências epidemiológicas em unidades interpretáveis, facilitando tanto consultas determinísticas quanto processos de interpretação baseados em linguagem. Como trabalhos futuros, pretende-se integrar as visões semânticas materializadas com ontologias de aplicação e grafos de conhecimento, permitindo validação semântica, interoperabilidade entre diferentes fontes de dados e maior explicabilidade no processo de recuperação. Também será investigado o uso de agentes de IA para automatizar a atualização incremental das visões semânticas materializadas à medida que novos microdados epidemiológicos forem disponibilizados.

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 e auxílio na elaboração das ilustrações técnicas.

Referências

- Arslan, M., Ghanem, H., Munawar, S., and Cruz, C. (2024). A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790. <https://doi.org/10.1016/j.procs.2024.09.178>.
- Biswal, A., Patel, L., Jha, S., Kamsetty, A., Liu, S., Gonzalez, J. E., Guestrin, C., and Zaharia, M. (2024). Text2sql is not enough: Unifying ai and databases with tag. <https://vldb.org/cidrdb/papers/2025/p11-biswal.pdf>.
- de Oliveira, R. B. D. and Alves-Souza, S. N. (2026). Aedestag: vigilância de arboviroses com documentos semanticamente estruturados e retrieval-augmented generation. In *I Workshop de Computação Aplicada às Doenças Tropicais Negligenciadas (CADTN) Evento integrante do 26º Simpósio Brasileiro de Computação Aplicada à Saúde (SB-CAS 2026)*.
- 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>.
- Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501. ACM. <https://doi.org/10.1145/3637528.3671470>.
- Fang, X., Xu, W., Tan, F. A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S., and Faloutsos, C. (2024). Large language models (llms) on tabular data: Prediction, generation, and understanding – a survey. <https://doi.org/10.48550/arXiv.2402.17944>.
- José, M. M., Cação, F. N., Ribeiro, M. F., Cheang, R. M., Pirozelli, P., and Cozman, F. G. (2025). Question answering with texts and tables through deep reinforcement learning. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 339–353, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-79032-4_24.
- Liu, T., Wang, F., and Chen, M. (2023). Rethinking tabular data understanding with large language models. <https://doi.org/10.48550/arXiv.2312.16702>.
- Long, L., Gu, X., Sun, X., Ye, W., Wang, H., Wu, S., Chen, G., and Zhao, J. (2025). Bridging the semantic gap between text and table: A case study on NL2SQL. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=qmsX2R19p9>.
- O’Shaughnessy, D. (2026). An overview of recent advances in natural language processing for information systems. *Applied Sciences*, 16:1122. <https://doi.org/10.3390/app16021122>.
- Yu, X., Jian, P., and Chen, C. (2025). Tablerag: A retrieval augmented generation framework for heterogeneous document reasoning. <https://doi.org/10.48550/arXiv.2506.10380>.