6º SIMPIF | SIMPÓSIO DE PESQUISA INOVAÇÃO E PÓS-GRADUAÇÃO DO INSTITUTO FEDERAL DA PARAÍBA
2025 23 a 25 julho
MAR DE IDEIAS
Ciência, Inovação e Sustentabilidade para um planeta em transformação

Arquitetura para gerar anotações semânticas de dados JSON de APIs com modelos de linguagem de larga escala

Klivio Rafael Nunes e Silva (IFPB, Campus João Pessoa), Diego Ernesto Rosa Pessoa (IFPB, Campus João Pessoa)

E-mails: klivio.silva@academico.ifpb.edu.br, diego.pessoa@ifpb.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

Palavras-chave: Anotação Semântica; Linguagem Natural; APIs; Dados Estruturados

1. Introdução

O recente crescimento no volume de dados publicado na Web traz à tona a demanda por mecanismos que garantam o seu enriquecimento semântico, permitindo a compreensão tanto por humanos, como por máquinas e também promovendo interoperabilidade e reuso de informações (BERNERS-LEE; LASSILA; HENDLER, 2001). Embora APIs que expõem dados JSON sejam amplamente utilizadas, frequentemente carecem de metadados descritivos que esclareçam o significado semântico de suas estruturas, dificultando a interpretação automatizada e o alinhamento de vocabulários (BANIAS *et al*., 2021).

Diante desse cenário, propõe-se uma arquitetura semiautomática que combine Modelos de Linguagem de Larga Escala (LLMs) com o vocabulário Schema.org, vocabulário colaborativo amplamente reconhecido por mecanismos de busca e aplicações. Utilizando seus tipos e propriedades, os dados se tornam mais conectados e compreensíveis por máquinas (SCHEMA.ORG, 2011). O objetivo é sugerir anotações semânticas para dados JSON de APIs, por meio da adição de metadados que explicitem seu significado com base em terminologias padronizadas. A estratégia aproveita a capacidade dos LLMs de associar pares chave-valor a propriedades semanticamente relevantes, com validação posterior realizada por um avaliador humano.

Este trabalho propõe uma arquitetura modular para a anotação semântica de dados JSON de APIs, fundamentada no uso de um vocabulário ontológico e na validação humana. Para validação da arquitetura, foi implementado um estudo de caso em que adotou-se o Schema.org como vocabulário semântico de referência, devido à sua ampla aceitação e cobertura de domínios, e o Ollama, uma plataforma open-source, para a execução local de LLMs, assegurando privacidade e independência de conexões externas. Esta implementação realiza todas as etapas previstas na arquitetura, resultando em um documento no formato JSON-LD, uma extensão do *JSON* que permite representar dados estruturados e vinculá-los à Web Semântica usando vocabulários.

2. Trabalhos Relacionados

O enriquecimento semântico de dados estruturados com LLMs tem crescido como tema de pesquisa nos últimos anos. De Mendonça e Arakaki (2025) discutem o uso de ferramentas para inserção de metadados em registros bibliográficos, com foco em estruturas já existentes em bases acadêmicas, mas sem aplicação de modelos de linguagem para geração automática de anotações. Freitas (2022) foca em anotação semântica baseada em ontologias para análise de conteúdo textual, não dados estruturados de APIs.

Estudos de Wei *et al*. (2022) e Min *et al*. (2023) mostram que LLMs têm boa capacidade de inferência semântica, mas apresentam dificuldade em seguir rigorosamente vocabulários controlados. Esses vocabulários são listas pré-definidas de termos, como propriedades específicas de uma ontologia. Por exemplo, um LLM pode entender que "cor do produto" se refere a cor, mas sugerir "*colorProduct*" ou "*itemColor*" em vez da propriedade exata do Schema.org, "*color*", gerando resultados semanticamente próximos, porém imprecisos lexicalmente.

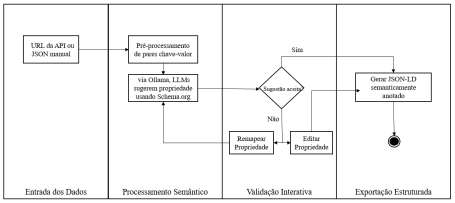
A originalidade deste trabalho reside, portanto, na proposta de uma arquitetura, que possibilite a execução do fluxo completo e validável para o mapeamento semântico de dados JSON de APIs. A atuação de um validador humano no processo garante a conformidade com o vocabulário e o contexto do dado.

3. Materiais e Métodos

3.1 Arquitetura proposta

A arquitetura proposta é composta por quatro módulos principais: entrada de dados, processamento semântico, validação interativa e exportação estruturada. O fluxo da comunicação entre os módulos é apresentado na Figura 1.

Figura 1. O fluxo de atividades da arquitetura proposta.



Fonte: Autores (2025).

No módulo de Entrada de Dados, o sistema permite importar dados JSON via URL de API externa ou inserção manual em formato de texto. Os dados brutos passam por um pré-processamento para garantir a consistência mínima necessária à interpretação pelos LLMs. Essa etapa inclui validação da estrutura JSON, identificação de objetos malformados, simplificação de aninhamentos e padronização dos pares chave-valor em formato uniforme.

O módulo de Processamento Semântico utiliza dois LLMs: o *LLaMA 3: latest* (Meta) devido ao seu bom processamento de linguagem e capacidade de seguir instruções complexas e o *DeepSeek-R1: free* que foi projetado para tarefas de raciocínio, exibindo uma melhor compreensão de estruturas, funcionando bem no mapeamento de dados a vocabulários ontológicos. Ambos são executados localmente via Ollama.

Os dados processados são enviados ao modelo via um *prompt* estruturado. O *template* utilizado é padronizado para cada par de chave-valor, o modelo é apresentado na Figura 2.

Figura 2**.** Template do *prompt* estruturado enviado aos LLMs para anotação semântica.

```
prompt = (
    f"Baseado no Schema.org, qual a propriedade mais adequada para a seguinte chave e valor?\n"
    f"Chave: {key}\n"
    f"Valor: {value} (tipo: {type(value).__name__})\n"
    f"Responda com apenas o nome da propriedade do Schema.org."
)
```

Fonte: Autores (2025).

No módulo de Validação Interativa, um validador humano analisa cada sugestão de anotação fornecida por LLMs. Essa etapa é essencial para garantir a adesão ao vocabulário e corrigir imprecisões dos modelos. O avaliador pode, via interface gráfica (Figura 3), aceitar, remapear ou modificar a sugestão para uma propriedade mais precisa. O esforço humano é minimizado, pois o sistema oferece uma sugestão inicial, transformando a anotação de uma criação do zero para uma validação e ajuste.

Figura 3**.** Interface de validação onde o usuário confirma ou ajusta propriedades sugeridas pelo modelo com base no Schema.org.

```
🔧 Chave: id
📌 Tipo: str
🔵 Propriedade Schema.org sugerida: identifier
Escolha uma opção - (1) Editar propriedade, (2) Mapear novamente, (3) Confirmar: 3
✔ Propriedade confirmada sem alterações.
```

Fonte: Autores (2025).

Por fim, no módulo de Exportação Estruturada, o sistema gera um documento no formato *JSON-LD*. As propriedades anotadas não modificam o *schema* original da *API*, mas são incorporadas como metadados descritivos em um arquivo anexo. É possível identificar a anotação, devido a dois campos *@context* e *@type*, os quais apontam para os vocabulários utilizados, no caso Schema.org e a propriedade semântica associada à chave-valor.

4. Resultados e Discussão

Para avaliação preliminar, foram utilizadas 300 amostras *JSON* anotadas manualmente. Cada amostra consiste em um par chave-valor (“*key”* e “*value”*) e a propriedade correspondente conforme o Schema.org (“*expect”*). A anotação manual foi realizada por especialista, e o número de amostras, com seus pares chave-valor individuais, foi definido como um conjunto representativo inicial para testes de viabilidade da arquitetura, cobrindo uma variedade de tipos de dados comuns em APIs. O formato utilizado para cada amostra de teste é apresentado na Figura 4.

Figura 4**.** Exemplo de uma amostra JSON utilizada no conjunto de dados de avaliação.

```
{
  "key": "data_category_info",
  "value": "Electronics",
  "expected": "category"
},
```

Fonte: Autores (2025).

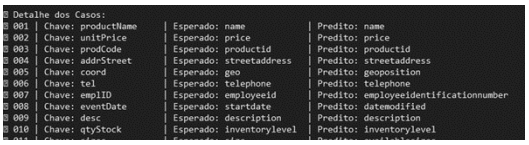
Cada sugestão do modelo foi comparada ao campo "*expected*" para cada par chave-valor da amostra. As métricas de Verdadeiros Positivos (VP) e Falsos Positivos (FP) foram calculadas por sugestão de par chave-valor: um VP ocorre quando a sugestão corresponde à propriedade esperada, e um FP, quando diferente.

O teste obteve um recall de 100%, indicando que o modelo sempre forneceu uma sugestão para cada par chave-valor, conforme esperado pela natureza generativa dos LLMs. Contudo, a precisão de 38% (114 VPs de 300) sinaliza a necessidade de refinamento e aprimoramento na capacidade de mapeamento dos LLMs ao vocabulário Schema.org.

Foi possível observar predições que eram lexicalmente semelhantes às propriedades corretas, mas fora do padrão esperado como, por exemplo, “*seller*” em vez de “*sellername*” ou “*geo*” por “*geoposition*”. Esse padrão de ação é consistente com o relatado por Bender *et al*. (2021) e Min *et al*. (2023), que observaram a dificuldade dos LLMs em

seguir vocabulários e ontologias, preferindo gerar termos semanticamente próximos, comportamento ilustrado na Figura 5.

Figura 5. Comparação entre propriedades esperadas e predições feitas pelo modelo LLM para os pares chave/valor. Observa-se uma alta taxa de cobertura, com pequenas variações lexicais nas predições incorretas.



Fonte: Autores (2025).

5. Considerações Finais

A arquitetura proposta mostrou-se tecnicamente viável para a tarefa de anotação semântica de dados *JSON* com uso de LLMs. A modularidade do sistema permite adaptações e melhorias futuras, como o refinamento de *prompts*, substituição de modelos e aplicação em outros domínios de dados.

Entretanto, os resultados iniciais apontam a necessidade de aprimoramento na precisão das anotações. Embora o sistema seja capaz de gerar sugestões, a qualidade ainda precisa ser calibrada. O sistema demonstra ser capaz de reduzir o esforço do validador humano ao fornecer uma sugestão base, servindo como ponto de partida para validação e ajuste, em vez de exigir que se anote do zero, desta forma otimizando a atividade.

Como trabalho futuro, pretende-se expandir o conjunto de dados anotados para treinamento e avaliação, aplicar técnicas de *fine-tuning* nos LLMs especificamente para o domínio do Schema.org e explorar outras abordagens de validação semântica assistida, para retreinar ou adaptar os modelos.

6. Referências

BANIAŞ, O.; FLOREA, D.; GYALAI, R.; CURIAC, D.-I. Automated Specification-Based Testing of REST APIs. **Sensors**, Basel, v. 21, n. 16, p. 5375, ago. 2021. Disponível em: https://www.mdpi.com/1424-8220/21/16/5375. Acesso em: 10 jul. 2025.

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT), 2021. **Proceedings...** New York: ACM, 2021. p. 610–623. Disponível em: https://doi.org/10.1145/3442188.3445922. Acesso em: 10 jul. 2025.

CONEGLIAN, C. S.; SEGUNDO, J. E. S. Inteligência artificial e ferramentas da web semântica aplicadas à recuperação da informação: um modelo conceitual com foco na linguagem natural. **Informação & Informação**, Londrina, v. 27, n. 1, p. 625–651, 2022. DOI: https://doi.org/10.5433/1981-8920.2022v27n1p625.

DE MENDONÇA, A. C. N.; ARAKAKI, A. C. S. Produção científica sobre enriquecimento semântico de metadados em dados bibliográficos utilizando a ferramenta Open Knowledge Maps. **Ciência da Informação Express**, [s. l.], v. 6, p. 1–21, 2025. DOI: https://doi.org/10.60144/v6i.2025.135.

DIJCK, J. van; POELL, T.; DE WAAL, M. **The Platform Society**: Public Values in a Connective World. Oxford: Oxford University Press, 2018. DOI: http://dx.doi.org/10.23860/MGDR-2018-03-03-08.

FREITAS, R. de. **Anotação semântica baseada em ontologia para análise de entrevistas dos atletas olímpicos brasileiros**. 2022. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-01122022-114642/publico/RovilsondeFreitas_revisada.pdf. Acesso em: 10 jul. 2025.

LASSILA, O.; HENDLER, J.; BERNERS-LEE, T. The semantic web. **Scientific American**, [s. l.], v. 284, n. 5, p. 34-43, 2001. DOI: doi:10.1038/scientificamerican052001-yL7Vw7HIOZ4iSjlnEeVsJ.

MIN, S.; LEWIS, M.; ZETTLEMOYER, L.; HAVRYLOV, S. **GPT-4 Doesn't Know It's Wrong**: An Analysis of Lexical Hallucination in Large Language Models. 2023. Disponível em: https://doi.org/10.48550/arXiv.2305.15352. Acesso em: 10 jul. 2025.

SCHEMA.ORG. *Schema.org*. [S.l.]: [s.n.], 2011–. Disponível em: https://schema.org/. Acesso em: 10 jul. 2025.

TAKAHASHI, T. (Org.). **Sociedade da informação no Brasil**: livro verde. Brasília, DF: Ministério da Ciência e Tecnologia, 2000. Disponível em: https://repositorio.mcti.gov.br/handle/mctic/650. Acesso em: 10 jul. 2025.

WEI, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022. Disponível em: https://doi.org/10.48550/arXiv.2201.11903. Acesso em: 10 jul. 2025.