LLMs na Geração de Mapeamentos Semânticos: Experimento no domínio de Plantas Industria

Geiza M. H. da Silva [2], Elvismary M. Armas[1], Pedro Saieg.[1], Rafael A. de Castro[1], Thiago D. Coqueiro[1], Melissa Lemos[1], Liester C. Castro[1]

[1]Instituto Tecgraf – Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO)
Rua Marquês de São Vicente, 225 – Gávea – RJ – Brazil

[2]Departamento de Informatica – Universidade Federal do Estado do Rio de Janeiro, UNIRIO
Av. Pasteur, 296 – Botafogo – RJ – Brazil

{emolina,pedrosf,rayrescastro,thiagodamicoc,melissa,liester}@tecgraf.puc-rio.br,
geiza.hamazaki@uniriotec.br

Abstract. *The application of Knowledge Graphs (KGs) has gained prominence in the implementation of solutions in several areas. In the Oil and Gas industry, its use is of interest due to the heterogeneity of the data present in the software supporting industrial plants. As standardization resources for this industry, Reference Data Libraries made available by international organizations and the use of the Industrial Data Ontology (IDO) stand out. However, applying these standards requires extending the IDO with the mapping of domain-specific concepts. In this context, this work proposes the study of the use of Large Scale Language Models (LLMs) as an assistant aiding understanding and suggesting semantic mappings.*

Resumo. *A aplicação de Grafos de Conhecimento (KGs) tem ganhado destaque na implementação de soluções em diversas áreas. Na indústria de Óleo e Gás, seu uso é de interesse devido à heterogeneidade dos dados presentes nos softwares de apoio a plantas industrias. Como recursos de padronização desta industria, destaca-se Bibliotecas de dados de Referências disponibilizadas por organizações internacionais e o uso da Ontologia de Dados Industriais (IDO). No entanto, aplicar esses padrões exige estender a IDO com o mapeamento de conceitos específicos do domínio. Neste contexto, este trabalho propõe o estudo sobre o uso de Modelos de Linguagem de Grande Escala (LLMs) como assistente auxiliando a compreensão e sugerindo mapeamentos semânticos.*

1. Introdução

Nos últimos anos, a aplicação de Grafos de Conhecimento (KGs na sua siglas em inglês) têm ganhado destaque na implementação de soluções em diversas áreas. Algumas soluções computacionais que apoiam setores industriais, incorporaram essa abordagem com o intuito de desenvolver ferramentas capazes de acessar, de forma semântica, o conhecimento dos especialistas do domínio.

Atualmente, a maioria dos *softwares* utilizados pelas industrias são especializados para um setor de uma empresa, não abrangendo todo o domínio do negócio. Entretanto, cada ferramenta possui seu próprio modelo de dados e terminologias, resultando em uma

visão fragmentada dos ativos industriais. A solução para uma visão unificada e consolidada dos ativos é a utilização de integração dos dados, e neste contexto os KGs oferecem uma estrutura para resolver esse problema, ao permitir a criação de uma camada semântica comum, integrando dados de diferentes fontes e promovendo uma compreensão unificada dos sistemas, possibilitando a descoberta de conhecimento.

O uso de Grafos de Conhecimento na indústria de Óleo e Gás tem sido investigado, pelos principais orgãos de desenvolvimento de padrões para troca de dados (CFIHOS[1], DEXPI[2], PCA[3]) evidenciando seu potencial para estruturar os dados gerados em todas as fases do ciclo de vida de uma planta industrial, potencializando ganhos econômicos e avanços em sustentabilidade ambiental. Um dos principais esforços nesse sentido é o desenvolvimento e a utilização da Ontologia de Dados Industriais (IDO), que busca promover a integração de dados durante todo o ciclo de vida das plantas industriais. Outras iniciativa, como a *Capital Facilities Information Handover Specification(CFIHOS)*, propõe uma padronização das informações necessárias para a transferência de dados entre as partes envolvidas nos projetos industriais. Entretanto, para que esses padrões sejam efetivamente utilizados, é necessário realizar o mapeamento dos conceitos específicos do domínio (como equipamentos, documentos ou processos) para os termos definidos na ontologia/KG, o que exige conhecimento técnico e semântico. Diante dessa dificuldade, a proposta assistente nesse processo, aproveitando sua capacidade de compreensão contextual e geração de sugestões coerentes para acelerar e facilitar o mapeamento semântico entre os domínios específicos e as ontologias/KGs.

2. Background

2.1. Engenharia digital e seus dados

A gestão da informação é parte fundamental na organização de uma empresa, sendo necessário, independente do contexto o estabelecimento de diretrizes e em alguns casos especificações técnicas. No domínio de plantas industriais, isso se torna imprescindível dado a necessidade de integração/interoperação entre os sistemas. Entretanto, por mais que existam nas empresas diretrizes e especificações técnicas definidas, tanto os requisitos de dados (para cada classe de componente dentro de um ativo de produção (ex. plataforma)) bem como o formato de armazenamento destes, a prática destes apresenta desafios. É comum, em projetos de grande escala, que a empresa contratante subdivida o escopo de trabalho em duas ou mais empresas contratadas, seja para mitigar riscos e/ou para acelerar o processo de entrega do projeto, uma vez que não existem tantas empresas com o porte/expertise necessários para a execução do projeto como um todo. Além disso, estas contratadas, em alguns casos, subcontratam parte da execução de um projeto. Este cenário dificulta a fiscalização por parte da contratante, que acaba direcionando a maior parte da energia em garantir "apenas" a completude necessária para que o projeto seja executado corretamente e não tanto para a qualidade da informação entregue.

Vale salientar a existência da possibilidade de que as diferentes empresas contratadas utilizem a mesma ferramenta com modelagens distintas do Banco de Dados, ou ferramentas diferentes, com estruturas de armazenamento de dados próprias, uma vez que

[1]CFIHOS:`https://www.jip36-cfihos.org`, acessado 27/06/2025.
[2]DEXPI: `https://dexpi.org/`, acessado 27/06/2025.
[3]PCA: `https://rds.posccaesar.org/WD_IDO.pdf`, acessado 27/06/2025.

a contratante costuma ter como maior preocupação a viabilidade da execução do projeto, e não como as informações são geradas, o que impacta a gestão futura do ativo, pois estes dados serão utilizados durante o ciclo de vida do ativo (operação, manutenção e descomissionamento).

2.2. Grafo de conhecimento e Modelos de Linguagem de Grande Escala

Nas duas ultimas décadas, os Grafos de Conhecimento (KGs) têm sido amplamente estudados, dando origem a diversas descrições e definições. Em Ehrlinger et al. [Ehrlinger and Wöß 2016] um grafo de conhecimento e caracterizado como: "um grafo de conhecimento adquire e integra informações em uma ontologia possibilitando a aplicação de raciocínio para derivar novos conhecimentos". Um KG pode ser entendido, de forma concisa, como "um grafo de dados destinado a representar conhecimento"[Bonatti et al. 2019].

Sendo uma estrutura de representação do conhecimento, onde os nós correspondem a conceitos que descrevem o domínio em questão, enquanto as arestas expressam as relações entre esses conceitos; o KG permite formalizar a semântica dos termos utilizados, possibilitando o acesso à informação explícita e a realização de inferências — via consultas sobre conhecimento explícito ou implícito — além da identificação de possíveis inconsistências [Bonatti et al. 2019].

O uso de KGs na indústria de Óleo e Gás tem sido explorado por alguns estudos (por exemplo: [Brewton 2023] e [Huang et al. 2020]), destacando seu potencial para organizar e estruturar os dados gerados ao longo do ciclo de vida de uma planta industrial. Essa abordagem permite apoiar a tomada de decisões estratégicas, com impactos positivos tanto no desempenho econômico quanto na sustentabilidade ambiental das operações.

Por outro lado, Modelos de Linguagem de Grande Escala (LLMs na sua sigla em inglês) têm ganhado destaque por seu uso em uma ampla gama de tarefas. Estudos recentes mostram que LLMs podem acelerar significativamente o processo de pesquisa ao automatizar tarefas repetitivas e ao oferecer suporte inteligente à análise e integração de dados complexos [Moor et al. 2023, Korinek 2023, Zheng et al. 2025].

Estudos recentes têm explorado maneiras de integrar os LLM à construção automatizada de grafos de conhecimento. No entanto, ainda não há um consenso consolidado sobre qual é a metodologia mais eficiente, robusta e escalável para essa tarefa. A seguir, é apresentado algumas pesquisas sobre geração de Grafos de conhecimento com LLM. [Sequeda et al. 2025] defende que o uso de KGs é essencial para garantir a confiança nas respostas fornecidas por sistemas baseados em LLMs. Os autores demonstram que os KGs não apenas melhoram a acurácia dos modelos, reduzindo significativamente erros semânticos e alucinações, como também oferecem mecanismos formais para explicabilidade e governança dos dados. Em seus experimentos, demonstraram que sistemas que combinam LLMs com KGs chegam a ter uma melhoria de até 4 vezes na precisão das respostas quando comparados com abordagens que usam apenas os esquemas SQL tradicionais. Além disso, o trabalho destaca que os LLMs podem simplificar os processos de engenharia de conhecimento, por meio de abordagens no estilo co-piloto.

[Song et al. 2024] apresenta um processo de geração de grafos de conhecimento começando com o *Document Chunking*, dividindo o texto em blocos menores mantendo o contexto. Em seguida, LLMs extraem e classificam entidades, que são conectadas por

similaridade (*Entity Linking*) e hierarquia (*Relationship Linking*). Por fim, formam-se grupos de entidades organizadas hierarquicamente. O grafo é usado conjuntamente com LLM para recuperação de informações. Durante a fase de recuperação, o LLM consulta o grupo de entidades, que foram construídas com base nas consultas dos usuários, permitindo a geração de respostas precisas a consultas. Esta abordagem foi testada com textos turísticos (blogs, redes sociais) gerando relatórios em linguagem natural sobre o grafo gerado.

Na pesquisa apresentada em [Carta et al. 2024] é utilizado textos de domínio específico (culinária, biomedicina) onde são extraidas as palavras-chave de cada documento (*Keywords extraction*) usando a LLM da Google *KeyphraseTransformer* e, a partir delas, geram-se hiperônimos que representam tópicos (*Topic Discovery*) usando a LLM Zephyr para generalizar essas palavras em tópicos mais amplos. Esses tópicos são agrupados em *clusters* relacionados usando técnicas de *embeddings* e o algoritmo de *Hierarchical Clustering* (HCA). Finalmente, cada um é rotulado e descrito para facilitar a compreensão, utilizando técnicas de prompt para LLM com o Zephyr.

Em [Cao et al. 2024] é utilizado textos médicos como dados de entrada e gera como saída um grafo, contendo triplas e nós padronizados utilizando o Neo4G. O processo envolve a divisão dos dados devido à limitação de tokens dos LLMs. Em seguida, realiza-se a extração e classificação de entidades via *prompts* e *string matching*, e a extração de relações entre entidades médicas utilizando LLMs e o contexto ontológico. A etapa de calibração de entidades assegura que apenas informações relevantes sejam incluídas no grafo. Por fim, remove-se a duplicidade de nós e resolvem-se anáforas para construir um grafo limpo e navegável.

O trabalho [Cremaschi et al. 2025] apresenta contribuições importantes para a tarefa de Anotação de Entidades em Células (CEA) no contexto da Interpretação Semântica de Tabelas (STI). Primeiramente, foi criado um *Gold Standard* com um conjunto de dados composto por tabelas com anotações detalhadas e informações semânticas estruturadas, voltadas à desambiguação de entidades — esse conjunto também pode ser utilizado na geração de *prompts* para ajuste fino ou testes de modelos de linguagem. Além disso, é realizada uma análise abrangente de diferentes estratégias de *prompt engineering*, com o objetivo de identificar estruturas que maximizem a qualidade das anotações. Por fim, o estudo inclui o *fine-tuning* do modelo Mixtral 8x7B especificamente para a tarefa de CEA, aprimorando sua capacidade de realizar anotações semânticas precisas em tabelas.

[Tupayachi et al. 2024] propõem uma metodologia para criação de ontologias. O design conceitual do fluxo de trabalho autônomo é composto por quatro módulos principais: (a) Aquisição de Fontes de Conhecimento, (b) Pré-processamento das Fontes de Conhecimento, (c) Geração de Ontologias com LLM, e (d) Implementação da Ontologia. O trabalho utiliza como entrada PDFs científicos para gerar uma ontologia em formato OWL. O processo inicia com a busca por artigos, seguida do pré-processamento com remoção de ruídos, separação de conteúdo e aplicação de processamento de linguagem natural em texto e imagem. Em seguida, um LLM identifica entidades e as organiza hierarquicamente em classes, propriedades e indivíduos, passando por etapas como remoção de *stop words*, normalização, detecção de sentenças e tokenização. Parte do pré-processamento é feito externamente (com NLTK) e parte é integrada à API do ChatGPT, antes de converter a ontologia gerada de JSON para OWL.

As metodologias analisadas, embora distintas em seus objetivos e domínios, apresentam pontos de interseção significativos que podem ser sumarizados na Tabela 1:

Tabela 1. Comparação entre abordagens de geração de conhecimento com LLMs

Critério	[Song et al. 2024]	[Carta et al. 2024]	[Cao et al. 2024]	[Cremaschi et al. 2025]	[Tupayachi et al. 2024]
Extração de Tabelas vs Texto	Texto	Texto	Texto	Tabelas convertidas	Tabelas convertidas
Hierarquização Semântica	Sim (tópicos)	Sim (tópicos)	—	—	Sim (ontologia)
Uso de Ontologias Externas vs. Geração Livre	Geração livre	Geração livre	Ontologias especializadas	Wikidata	Ontologias especializadas
Geração e Estruturação de Triplas	Não	Não	Gera triplas	Conversão indireta	Gera triplas
Validação Semântica e Controle de Qualidade	Não	Não	Sim (calibração)	Sim (fine-tuning)	Sim (lógica)
Formatos de Saída e Interoperabilidade	Saída interna	Saída interna	Neo4j	JSON conversível	RDF / OWL

2.3. Padronização da Representação de dados no setor de Óleo e Gás

O Projeto *Requirement Asset Digital Life-cycle Information*(READI)[4] foi uma iniciativa de digitalização no setor de Óleo e Gás, com o objetivo de padronizar e automatizar processos críticos de negócio relacionados a manutenção, operações, descarte de ativos e transformação de processos empresariais, além de desenvolver uma linguagem digital comum e uma estrutura unificada. Um dos resultados do projeto foi a contribuição com a construção da Ontologia de Dados Industriais (IDO) [5]. Esta é uma ontologia de nível superior, resultante da evolução da ISO 15926 - Parte 14 e derivação da *Basic Formal Ontology*(BFO). A IDO foi projetada para ser utilizada em todas as fases do ciclo de vida de ativos e processos industriais, servindo como base para a construção de vocabulários e para o gerenciamento de modelos de ativos que utilizam Bibliotecas de Dados de Referência (RDLs). A IDO define termos genéricos aplicáveis a diferentes domínios industriais, apresentada em uma estrutura modular e flexível, ela facilita a integração de dados provenientes de diversas fontes e sistemas ao longo do ciclo de vida do ativo. Esta ontologia constitui a parte inicial da nova norma ISO/NP 23726 *Ontology Based Interoperability*[6].

3. Experimentos e Resultados

Com conhecimento dos padrões de dados para troca de Dados no domínio de Óleo e Gás, das pesquisas relacionadas a utilização de LLMs da geração de Grafos de Conhecimentos(KGs) e com questionamentos de quão fácil seria utilizar um LLM para criar KGs passando como entrada um conjunto de arquivos, foi decidido explorar a utilização de LLM, através da técnica de *prompt* para a criação de mapeamentos semânticos, como fase inicial para construção de um KG, para alguns conceitos comuns a todas plantas industriais. O primeiro passo foi a escolha das entidades e atributos a serem mapeados. Esta seleção foi realizada a partir do conhecimento de engenheiros especialistas que lidam com a análise de documentos e consultas a bancos de dados de diferentes ferramentas. Estes especialistas dão apoio às equipes de diversas fases do ciclo de vida da indústria de óleo de gás e múltiplas plataformas. Foram escolhidos 20 conceitos e atributos, comuns a

[4]READI: `https://readi-jip.org.`, acessado 27/06/2025.

[5]PCA: `https://rds.posccaesar.org/WD_IDO.pdf`, acessado 27/06/2025.

[6]Está disponível em: `https://rds.posccaesar.org/WD_IDO.pdf`, acessado 27/06/2025.

qualquer projeto de planta industrial, para a análise inicial do estudo de caso. Seguem os conceitos selecionados e seus relacionamentos: 1) Piping relaciona com Pipeline; 2) Pipeline relaciona com: PipeRun e Fluid System; 3) PipeRun relaciona com Pipe; 4) Pipe relaciona com: Insulation, Design Max Pressure, Design Max Temperature, Operating Max Pressure, Operating Max Temperature; 5) Insulation, com relacionamentos com Material, Temperature,Thickness; e 6) Equipment com relacionamentos com Insulation, Design Max Pressure, Design Max Temperature.

Em seguida, foram realizados experimentos para utilização de LLM como assistente para a criação de mapeameamentos semânticos extendendo a ontologia de nivel superior IDO, ou suas extensões, foram utilizadas como entrada nos *prompts*. No primeiro experimento, foi solicitado ao LLM o melhor mapeamento para o conceito "Piping". No Listing 1 é apresentado a requisição realizada.

Listing 1. Prompt 1.

```
Dada a ontologia de fundamentação industrial IDO no arquivo
   abaixo, responda à pergunta:
Arquivo: {LIST_IDO.rdf}
Pergunta: Qual seria o melhor mapeamento para o conceito de "
   Piping" no contexto de plantas industriais.
```

A resposta do Prompt 1 está disponível no anexo: `https://drive.google.com/drive/folders/1gG-bz79H6wRloI--pah73TYysKaVR9g8?usp=sharing` . A resposta apresentada pela LLM é bem explicativa, foram indicadas as possibilidades de classificação como: *PhysicalObject*, *InanimatePhysicalObject*, *System*, *FunctionalObject*, *hasFunctionalPart* ou *Stream*. As várias opções pode dificultar a decisão de qual o melhor mapeamento, além do fato que nem todas as possibilidades são corretas, por exemplo a indicação de classificação como subclasse de *Stream* ou como subpropriedade de *hasFunctionalObject*. Mas a classificação de Piping de acordo com especialista do domínio é como *System*.

Como segundo experimento foi pedido para o LLM classificar os conceitos "Pipeline", "Pipeline Fluid System" e "Piperun" (Listing 2) especificando "Piping" é um Sistema. Além de fornecer um modelo relacionando "Pipeline", "Pipeline Fluid System" e "Piperun".

Listing 2. Prompt 2.

```
Dada a ontologia de fundamentação industrial IDO no arquivo
   abaixo, e que o conceito Piping é mapeado como um Sistema,
   responda à pergunta:
Arquivo: {LIST_IDO.rdf}
Pergunta: Qual a classificação para Pipeline , Pipeline Fluid
   System e Piperun e forneça um modelo na IDO relacionando
   Pipeline Fluid System, Pipeline e PipeRun
```

A resposta para o Prompt 2 está disponível em `https://drive.google.com/drive/folders/1gG-bz79H6wRloI--pah73TYysKaVR9g8?usp=sharing`. As definições apresentadas na resposta pelo LLM são corretas, mas restritivas

a classificação proposta de Pipeline como *System*. Diferentemente do que foi respondido no (Listing. 1) onde para um conceito foram apresentadas algumas classificações válidas; neste caso ”Pipeline” e ”Piperun” foram classificados como *System* e *FunctionalObject*, respectivamente, sendo que poderiam ser modelados também como um *PhysicalObject*, pois depende da visão que quer se ter no ativo. O LLM reconheceu que ”Pipeline” é um conceito relacionado com ”Pipeline Fluid System”e ”Piperun”.

Aprimorando os resultados anteriores, foi realizado o terceiro teste onde foi solicitado para classificar todos os conceitos (Listing3) e retornar o resultado em forma tabular (Tabela2). O arquivo de entrada fornecido (Listing 3) está disponível em: `https://drive.google.com/drive/folders/1gG-bz79H6wRloI--pah73TYysKaVR9g8?usp=sharing`.

Listing 3. Prompt 3.

```
Dada a ontologia de fundamentação industrial IDO no arquivo
   abaixo, já estendida com os conceitos Piping, Pipeline Fluid
   System e Piperun, responda à pergunta:
Arquivo: {IDO_EXTENDED_V1.RDF}
Pergunta: quais seriam as classes usadas para mapear as
   seguintes entidades:
  PipeRun, PipeRun Design Max Pressure, PipeRun Design Max
     Temperature, Insulation Material,
  Insulation Temperature, Insulation Thickness, PipeRun Operating
      Max Pressure, PipeRun Operating Max Temperature, Weight Dry
     , Weight Wet.
Gere a resposta em formato csv, primeira coluna o nome da
   entidade, segunda coluna o nome do conceito mapeado na
   ontologia IDO.
```

Tabela 2. Mapeamento de Entidades para Conceitos da Ontologia IDO geradas pelo LLM

Entidade	**Conceito Mapeado na Ontologia IDO**
Pipe	Pipeline
PipeRun Design Max Pressure	ScalarQuantityDatum
PipeRun Design Max Temperature	ScalarQuantityDatum
Insulation Material	MaterialCompositionQuality
Insulation Temperature	ScalarQuantityDatum
Insulation Thickness	ScalarQuantityDatum
PipeRun Operating Max Pressure	ScalarQuantityDatum
PipeRun Operating Max Temperature	ScalarQuantityDatum
Weight Dry	ScalarQuantityDatum
Weight Wet	ScalarQuantityDatum

Analisando a classificação proposta, tem-se que a LLM propôs ”Pipeline” como subclasse de *System*, mas ”Pipe” é um *PhysicalObject*. As outras entidades não estão

classificadas de maneira incorreta, mas a modelagem é muito descritiva, pois por exemplo ”PipeRun Design Max Pressure”, além de ter um número e uma unidade, ele trata de um adjetivo relativo a ”Design Pressure” que é a qualidade de ser a pressão máxima de projeto. O modelo atende a consulta para acesso explícito ao dado, mas se for realizado consultas relacionadas a extração de conhecimento implícito esta modelagem é muito restrita.

Como último experimento Listing4 foi solicitado para o LLM completar a ontologia para o conceito de ”Equipment” e suas propriedades, a partir do exemplo de ”Pipe”. A Fig. 1 mostra como o LLM conseguiu gerar o conceito de ”Equipment” e as propriedades associadas a ele usando a propriedade *hasQuality*, mas o LLM não integra os novos conceitos na ontologia passada como entrada, pois nela as entidades ”Design Max Pressure” e ”Design Max Temperature”, já estavam modeladas e não estão classificadas como *Thing*.

Listing 4. Prompt 4.

```
Dada a ontologia de dominio que ja estende a ontologia de
   fundamentação industrial IDO passada no contexto, responda à
   pergunta:
Arquivo: {arquivo rdf}
Pergunta: "contexto": contexto,
   "input":''complete a ontologia entrada colocando a entidade:
  Equipment
  com suas propriedades:
  Design Max Pressure
  Design Max Temperature

  Verifique como foi feito para o conceito Pipe. Gere a resposta
     em um arquivo rdf.
```

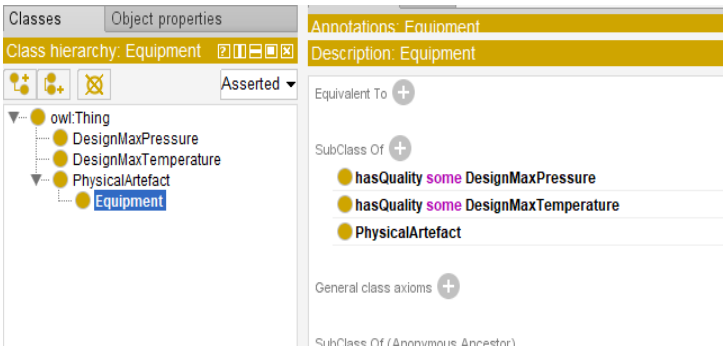

Figura 1. Resposta da LLM para conceito de Equipment, de forma assistida.

Os experimentos foram conduzidos por meio de chamadas à API da OpenAI, disponibilizada na infraestrutura em nuvem da Microsoft Azure, com a utilização do modelo GPT-4o. Para a implementação, recorreu-se à linguagem de programação Python e à biblioteca LangChain.

Os *prompts* apresentados neste trabalho é o resultado de algumas tentativas de utilização da LLM para realizar a proposta do grafo de conhecimento. Inicialmente, as

respostas não eram razoáveis para um grafo de conhecimento do domínio de óleo e gás.

Para comparação dos modelos foi criado um modelo semântico resultante da extensão manual da ontologia IDO com os conceitos da engenharia selecionados, usando a linguagem RDF/OWL e o software Protégé[7]. O mapeamento semântico resultante se encontra disponível em `https://drive.google.com/drive/folders/1gG-bz79H6wRloI--pah73TYysKaVR9g8?usp=sharing`. Este grafo foi estendido de forma incremental para poder se ajustar aos testes realizados com o LLM, representando o modelo com uma estrutura hierárquica com relações entre os conceitos que reflete melhor a semântica em um contexto, não sendo analisado somente o tipo do dado o que resulta uma modelagem mais plana. Vale salientar que esta modelagem segue o que é preconizado pelos principais órgãos de desenvolvimento de padrões para troca de dado no domínio de óleo e gás.

4. Conclusões

Dado o cenário de várias pesquisas sobre geração de Grafos de conhecimento (KGs) com LLM, neste trabalho foi apresentado experimentos com objetivo de explorar o uso de LLMs como assistente na criação modelos semânticos, aplicado ao domínio de plantas industrias, utilizando a Industrial Data Ontolgy (IDO) como ontologia superior. Após a análise nos testes realizados na seção 3.3, os LLMs apresentam-se como uma ferramenta útil para auxiliar tanto um ontologista quanto um especialista no domínio no processo de classificação de um conceito, mas com a ressalva de sempre solicitar a reposta com todas as possíveis classificações e analisar as respostas dado pode ser sugeridos modelos que não descrevem semanticamente o que é desejado pelo especialista. Foi verificado que o uso de LLM sem informação extra, não foi capaz de reconhecer os possíveis conceitos e relacionamentos entre eles no processo de construção do modelo semântico. Entretando, fornecendo um arquivo com um conceito semelhante modelado, e fazendo a indicação do conceito existente, a resposta foi satisfatória apesar de não retornar um arquivo com a extensão da ontologia fornecido como entrada. A ontologia resultante do uso de LLM possui uma estrutura plana e com foco no tipo de dados, reduzindo o potencial semântico de representação do conhecimento sobre o domínio e provavelmente isso dificultará a extração de conhecimento implícito do modelo gerado. O modelo proposto difere do preconizado pelas organizações internacionais relacionados a troca de informações no dominio industrial, principalmente óleo e gás(CFIHOS,PCA e DEXPI).

Como trabalhos futuros em relação a utilização de LLMs para gerar KGs tem-se :

- Inserir a Biblioteca dos dados de referência e exemplos dos modelos semânticos fornecidos pelas organizações internacionais e verificar a geração dos KGs;
- Validar a utilização de metodologias propostas ou propor uma metodologia para utilizar LLMS para gerar ontologias e KGs;
- Verificar como se comporta a criação dos KGs relacionados com a visão sobre a planta industrial (Função, Localização e Instalação); e
- Verificar a possibilidade de propor conjunto de *prompts* para ser customizado para criação modelos semânticos relacionados a área de plantas industriais.

[7]Protégé: `https://protege.stanford.edu`, acessado 27/06/2025.

Referências

Bonatti, P. A., Decker, S., Polleres, A., and Presutti, V. (2019). Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web (Dagstuhl Seminar 18371). *Dagstuhl Reports*, 8(9):29–111.

Brewton, B. (2023). How Using Knowledge Graphs can Optimize the Oil and Gas Industry. https://www.linkedin.com/pulse/how-using-knowledge-graphs-can-optimize-oil-gas-industry\-jon-brewton#. Acessado em 27/06/2025.

Cao, L., Sun, J., and Cross, A. (2024). An automatic and end-to-end system for rare disease knowledge graph construction based on ontologies-enhanced large language models. https://arxiv.org/abs/2403.00953. Acessado em 27/06/2025.

Carta, S., Giuliani, A., Manca, M. M., Piano, L., and Tiddia, S. G. (2024). Towards zero-shot knowledge graph building: Automated schema inference. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, page 467–473. Association for Computing Machinery.

Cremaschi, M., D'Adda, F., and Maurino, A. (2025). steellm: An llm for generating semantic annotations of tabular data. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. In Martin, M., Cuquet, M., and Folmer, E., editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016*, volume 1695 of *CEUR Workshop Proceedings*.

Huang, S., Wang, Y., and Yu, X. (2020). Design and implementation of oil and gas information on intelligent search engine based on knowledge graph. *Journal of Physics: Conference Series*, 1621:012010.

Korinek, A. (2023). Language models and cognitive automation for economic research. Working Paper 30957, National Bureau of Economic Research. Acessado em 27/06/2025.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Sequeda, J., Allemang, D., and Jacob, B. (2025). Knowledge graphs as a source of trust for llm-powered enterprise question answering. *Journal of Web Semantics*, 85:100858.

Song, S., Yang, C., Xu, L., Shang, H., Li, Z., and Chang, Y. (2024). Travelrag: A tourist attraction retrieval framework based on multi-layer knowledge graph. *ISPRS International Journal of Geo-Information*, 13(11).

Tupayachi, J., Xu, H., Omitaomu, O. A., Camur, M. C., Sharmin, A., and Li, X. (2024). Towards next-generation urban decision support systems through ai-powered construction of scientific ontology using large language models—a case in optimizing intermodal freight transportation. *Smart Cities*, 7(5):2392–2421.

Zheng, T., Deng, Z., Tsang, H. T., Wang, W., Bai, J., Wang, Z., and Song, Y. (2025). From automation to autonomy: A survey on large language models in scientific discovery. https://arxiv.org/abs/2505.13259. Acessado em 27/06/2025.