

Uma proposta de Modelo Arquitetural Neuro-Simbólico para suporte à avaliação de Políticas de CT&I baseada em evidências

Ricardo Luiz de Carvalho Barbosa¹, Ricardo André Cavalcante de Souza²

¹Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

²Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Departamento de Computação
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

{ricardo.luiz, ricardo.souza}@ufrpe.br

Abstract. *Public policy management in Science, Technology, and Innovation (STI) confronts the challenge of data fragmentation. This work proposes a neuro-symbolic architecture for the automated population of Knowledge Graphs (KGs) leveraging Large Language Models (LLMs). The architecture is structured into four functional components: hybrid ingestion of STI metrics, processing via entity and relationship extraction pipelines, semantic persistence in RDF, and application for decision support. Consequently, the solution transforms raw data into actionable intelligence, enabling the identification of innovation gaps and the optimization of evidence-based public resource allocation.*

Resumo. *A gestão de políticas públicas em Ciência, Tecnologia e Inovação (CT&I) enfrenta o desafio da fragmentação de dados. Este trabalho propõe uma arquitetura neuro-simbólica para o povoamento automatizado de Grafos de Conhecimento (KGs) utilizando Modelos de Linguagem de Larga Escala (LLMs). A arquitetura é estruturada em quatro componentes funcionais: ingestão híbrida de métricas de CT&I, processamento via pipelines de extração de entidades e relações, persistência semântica em RDF e aplicação para suporte à decisão. Como resultado, a solução transforma dados brutos em inteligência acionável, permitindo a identificação de lacunas de inovação e a otimização da alocação de recursos públicos baseada em evidências sólidas.*

1. Introdução

A gestão pública voltada às políticas de Ciência, Tecnologia e Inovação (CT&I) enfrenta atualmente um cenário de volume de dados crescente, caracterizado por uma alta dimensionalidade e volume de dados não estruturados que desafia as ferramentas convencionais de análise e monitoramento [Bergek et al. 2023]. Diariamente, o setor governamental é inundado por um volume massivo de dados provenientes de fontes heterogêneas e dinâmicas, incluindo currículos de pesquisadores, relatórios de fomento, depósitos de patentes e uma produção científica global em crescimento exponencial [Reese et al. 2021]. Esse fenômeno, embora rico em potencial, frequentemente resulta na formação de “ilhas de dados” em que a informação permanece isolada em silos departamentais, dificultando

a visão holística e transdisciplinar necessária para a condução estratégica de uma nação ou região [Arya et al. 2026].

A dificuldade central para a administração pública não reside apenas na coleta desses dados, mas na capacidade de conectar pontos dispersos para gerar políticas públicas baseadas em evidências sólidas [Reese et al. 2021, Arya et al. 2026]. Sem uma estrutura que relacione semanticamente o investimento público aos resultados de impacto, a tomada de decisão tende a ser fragmentada e reativa [Kalaycı et al. 2021, Bellomarini et al. 2022]. Conforme apontado por [Kalaycı et al. 2021], a gestão eficiente do conhecimento em paisagens informacionais densas exige mecanismos de exploração que permitam navegar por relações complexas, transformando dados brutos em conhecimento acionável para identificar lacunas de inovação.

Nesse contexto, os modelos tradicionais de análise falham ao não capturar a evolução dinâmica das redes de colaboração e a semântica oculta nos metadados de CT&I [Bellomarini et al. 2022]. A integração de métricas de CT&I via API exige uma infraestrutura que suporte tanto a escala quanto o rigor lógico. A emergência de arquiteturas neuro-simbólicas, que combinam o poder de extração de Grandes Modelos de Linguagem (LLMs, do inglês Large Language Models) com a estrutura de Grafos do Conhecimento (KG, do inglês Knowledge Graph), configura-se como sistemas híbridos que integram a expressividade linguística e a capacidade de inferência contextual dos modelos neurais (componente sub-simbólico) à semântica formal e ao rigor lógico das ontologias e grafos de conhecimento (componente simbólico). Essa união permite que o processamento estatístico de padrões seja mitigado por estruturas de conhecimento verificáveis e explicitamente relacionadas. Esse tipo de arquitetura oferece uma oportunidade única para automatizar a descoberta de padrões e enriquecer o grafo com novas métricas de impacto [Colombo et al. 2025]. Portanto, a proposta de um modelo conceitual que utilize KGs para unificar esses fluxos de informação apresenta-se como uma alternativa para modernizar a inteligência estratégica governamental e garantir a precisão analítica no ciclo de políticas públicas.

O objetivo deste trabalho é propor e formalizar uma arquitetura neuro-simbólica que utilize LLMs para a automação do povoamento de KGs, aprimorando a integridade semântica dos dados de CT&I através de restrições ontológicas, integrando dados de produtividade e impacto científico provenientes de uma API de métricas de CT&I.

A opção por uma arquitetura baseada em KGs, em substituição aos bancos de dados relacionais convencionais, justifica-se pela flexibilidade inerente às estruturas de grafos frente à rigidez dos esquemas tabulares. Enquanto modelos relacionais enfrentam limitações severas para escalar em cenários de alta conectividade e heterogeneidade de dados [Reese et al. 2021], os KGs permitem a modelagem fluida e a evolução contínua do esquema de dados sem a necessidade de reestruturações onerosas [Shimizu and Hitzler 2025].

Um dos pontos positivos desta proposta reside na capacidade de inferência e raciocínio lógico sobre o conhecimento estabelecido. Por meio do uso de ontologias formais, o modelo proposto visa permitir a descoberta de relações implícitas entre atores e tecnologias que não seriam capturadas por consultas SQL tradicionais. Essa capacidade de realizar buscas semânticas complexas transforma dados brutos em inteligência

estratégica, permitindo que o setor público antecipe tendências e otimize a alocação de recursos em CT&I com base em uma rede de conhecimento interconectada e autoexplicativa [Kalaycı et al. 2021, Bellomarini et al. 2022].

2. Fundamentação Teórica

Esta seção estabelece as bases conceituais e o estado da arte necessários para sustentar o modelo proposto, articulando as dimensões estratégica e tecnológica que compõem a solução. A fundamentação organiza-se em três eixos principais que se interconectam para viabilizar a inteligência estratégica em políticas de CT&I. Inicialmente, discute-se o papel da Gestão da Informação e do Conhecimento como suporte fundamental ao ciclo de políticas públicas e à tomada de decisão baseada em evidências. Em seguida, exploram-se os KGs e sua convergência com a Inteligência Artificial Generativa (GenAI), fundamentando-se na transição para arquiteturas neuro-simbólicas que garantem rigor lógico e automação na descoberta de conhecimento. Por fim, detalham-se os fundamentos das métricas de impacto científico (obtidos via API), definindo o domínio de dados que servirá de insumo para o enriquecimento semântico do modelo conceitual.

2.1. Gestão da Informação e do Conhecimento no Setor Público

A gestão de políticas públicas em CT&I enfrenta hoje o desafio do crescimento exponencial da literatura científica e dos registros de inovação, o que torna a curadoria manual e a anotação de dados inviáveis [Dessi et al. 2021]. Segundo [Malec et al. 2023], essa saturação informacional gera uma fragmentação da estrutura causal dos dados, impedindo que gestores identifiquem variáveis que conectam investimentos a resultados efetivos. Complementando essa visão, [Dreger et al. 2025] apontam que a persistência de formatos tabulares não padronizados agrava o problema, resultando em baixa reusabilidade e na criação de “pântanos de dados” (data swamps), em que a ausência de semântica inviabiliza a recuperação de conhecimento útil [Hoseini et al. 2024].

Para superar essa limitação da “memória institucional” humana, a governança de dados no setor público deve transitar de repositórios passivos para ecossistemas de dados ativos. [Meroño-Peñuela et al. 2025] argumentam que os KGs surgem como infraestrutura crítica para automatizar essa governança, permitindo o rastreamento semântico do ciclo de vida da informação. Essa transição exige o que [Abu-Salih et al. 2025] definem como interoperabilidade semântica: a capacidade de inferir contexto entre “ilhas de dados” díspares. Nesse cenário, a integração de Modelos de Linguagem de Grande Escala (LLMs) potencializa a harmonização de dados, auxiliando a curadoria humana intensiva por processos automatizados de extração de conhecimento computável [Abu-Salih et al. 2025, Malec et al. 2023].

De acordo com [Hoseini et al. 2024], a eficácia estratégica dessa abordagem reside na formalização do conhecimento de domínio em ontologias, o que permite que sistemas automatizados interpretem relações complexas antes restritas a especialistas. [Vidal et al. 2025, Wu et al. 2025] convergem ao afirmar que a utilização de grafos permite a justificação lógica das métricas de impacto; o gestor deixa de apenas visualizar dados isolados para compreender as evidências semânticas que sustentam determinada tendência de inovação.

Em última análise, a evolução da Gestão do Conhecimento em CT&I para modelos baseados em grafos viabiliza uma “visão holística” para a soberania estratégica [Ouzounis et al. 2023]. A capacidade de realizar o ranking de candidatos e projetos a partir de bases integradas (data-driven ranking) transforma o armazenamento bruto em inteligência acionável, garantindo transparência e responsabilidade na alocação de recursos públicos [Meroño-Peñuela et al. 2025, Ouzounis et al. 2023].

2.2. Grafos de Conhecimento e IA Generativa na Construção de KGs

A transição de sistemas de CT&I baseados em bancos relacionais para KGs justifica-se pela incapacidade dos modelos tradicionais em lidar com a heterogeneidade e conectividade dos dados científicos. Enquanto [Ouzounis et al. 2023] apontam que modelos relacionais falham na escalabilidade de dados não estruturados, a abordagem de grafos, enriquecida por Reconhecimento de Entidades Nomeadas (NER), permite converter patentes e relatórios em redes semânticas unificadas. Essa mudança de paradigma possibilita a tarefa de *Link Prediction*, em que algoritmos de aprendizado de máquina identificam relações ocultas entre investimentos e impactos científicos com alta acurácia [Ouzounis et al. 2023].

Todavia, o gargalo histórico para a viabilidade desses grafos reside na laboriosa criação manual de ontologias e mapeamentos. Para superar essa barreira, a literatura recente converge para o uso de GenAI e LLMs. [Hoseini et al. 2024, Dreger et al. 2025] demonstram que LLMs podem automatizar a interpretação de tabelas e a inferência de relacionamentos, tornando o processo agnóstico à estrutura da fonte original. Complementando essa visão, [Dessi et al. 2021] comprovam que sistemas híbridos, que combinam múltiplos extratores de Processamento de Linguagem Natural (PLN), superam abordagens isoladas ao maximizar a revocação (*recall*) na captura de triplas de conhecimento.

A fronteira atual dessa evolução é a arquitetura neuro-simbólica, que integra o poder indutivo dos LLMs ao rigor lógico dos grafos. [Benjira et al. 2025, Wu et al. 2025] argumentam que essa integração supera falhas de interoperabilidade semântica, utilizando técnicas de Engenharia de Prompt como *few-shot prompting* para extrair atributos granulares que consultas SQL tradicionais ignorariam. No entanto, a autonomia da GenAI não é absoluta: [Abu-Salih et al. 2025, Vidal et al. 2025] advertem sobre o risco de alucinações, propondo o uso de *Retrieval-Augmented Generation* (RAG) e linguagens de restrição como SHACL para garantir que o povoamento do grafo respeite as regras de negócio do domínio.

Estudos também apontam a importância da abordagem *Human-In-The-Loop* (HITL) em que especialistas da área atuam na verificação de resultados gerados por LLMs com o objetivo de validar tais informações e minimizar alucinações provenientes destes modelos [Doumanas et al. 2025].

Nesta arquitetura, o KG deixa de ser um mero repositório para tornar-se um “controlador” da IA. [Meroño-Peñuela et al. 2025] destacam que o grafo provê a estrutura necessária para governar o comportamento dos modelos generativos, garantindo a proveniência dos dados. Como defendido por [Malec et al. 2023], essa transparência lógica é o que permite transformar “ilhas de dados” em políticas baseadas em evidências, fornecendo a trilha justificável que conecta o investimento governamental ao resultado de

inovação.

2.3. Métricas de CT&I e Fonte de Dados

A visão de flexibilidade arquitetural, explanada por [Meroño-Peñuela et al. 2025] no framework KG.gov, que propõe uma modelagem em camadas (metadados, recursos e semântica) para assegurar a interoperabilidade e a integridade contextual de fontes heterogêneas.

De acordo com [Gokhberg et al. 2013], os indicadores de CT&I são classificados em seis categorias principais: Contexto, Insumos, Procedimento, Resultados, Ciclo de Desenvolvimento e Impacto (detalhados na Tabela 1).

Tabela 1. Classificação de Indicadores de CT&I

CATEGORIAS	TIPOS DE INDICADORES DE CT&I
Contexto	Ambiente econômico, social, político e ecológico, estruturas de incentivo.
Insumos	Recursos (financeiros, de pessoal, de conhecimento e tecnológicos) e financiamento público.
Procedimento	Colaboração, parcerias e transferência de tecnologia.
Resultados	Conhecimento tácito (graduados, diplomas e certificados), conhecimento codificado e tecnologias não incorporadas (publicações científicas, patentes e suas citações, licenças, know-how etc.), tecnologias incorporadas e inovações.
Ciclo de Desenvolvimento	Propriedade intelectual e comercialização de tecnologia, disseminação, utilização e consumo de tecnologia na indústria, no setor público e nos lares.
Impacto	Impactos diretos (vendas, emprego, adoção pelo utilizador final, etc.), impactos econômicos, sociais e outros impactos indiretos.

Nesse ecossistema, uma API de métricas de CT&I fornece os insumos necessários para superar a limitação dos metadados superficiais criticada por [Dessì et al. 2021]. Para esses autores, a utilidade de um grafo acadêmico reside em sua capacidade de ser “autoexplicativo”, o que, na arquitetura proposta, é viabilizado pela inclusão de métricas de impacto como pesos nas arestas do grafo [Ouzounis et al. 2023]. Ao associar indicadores de produtividade diretamente à topologia da rede, permite-se que o cálculo de impacto seja uma propriedade intrínseca do grafo, facilitando a identificação de lacunas de investimento de forma escalável [Benjira et al. 2025].

A viabilidade operacional do processo de ingestão depende, contudo, de automação e tratamento semântico. Enquanto [Reese et al. 2021] enfatizam o uso de arquivos de configuração (YAML) para a ingestão periódica e escalável de novas fontes, [Abu-Salih et al. 2025] introduzem o papel transformador dos LLMs na normalização de terminologias e extração de entidades. A sinergia entre a automação estrutural e a capacidade dos LLMs em lidar com ambiguidades reduz a necessidade de reestruturações manuais, permitindo que o KG evolua conforme os esquemas de dados de CT&I se transformam.

3. Modelo Conceitual Proposto

O modelo conceitual proposto fundamenta-se em uma arquitetura neuro-simbólica, projetada para mitigar o hiato entre a fluidez dos dados brutos de CT&I e o rigor semântico necessário à governança pública. A arquitetura é baseada em quatro componentes funcionais interdependentes, orquestradas para garantir a proveniência, a escalabilidade e a explicabilidade do conhecimento gerado.

3.1. Componentes Necessários para Implementação

A viabilização técnica do modelo proposto exige uma pilha tecnológica integrada (Figura 1). Os dados de métricas serão obtidos em tempo real a partir de: API's de métricas de CT&I (como a *SCOPUS* e *Web of Science*); e planilhas (.csv) oriundas de Sistemas de Informação de Pós-Graduação e Pesquisa, como a Plataforma Lattes (*CNPq*) e a Plataforma Sucupira (*CAPES*). Como componente central do componente de extração de relações e normatização, um LLM (como o *gemini-3-flash-preview*, *gpt-5.4*, *claude-sonnet-4-6* ou similar) atua como o motor de orquestração. Para garantir a interoperabilidade e a integridade semântica, a Ontologia de Domínio é formalmente modelada através de uma ferramenta de engenharia ontológica baseada em OML (como o Rosetta). O uso desse tipo de ferramenta permite a especificação rigorosa de classes e propriedades (Pesquisador, Instituição, Impacto e Fomento) sob o padrão OML (Ontological Modeling Language), assegurando que o vocabulário unificado de CT&I seja traduzido com precisão para representações OWL (o padrão OML permite definir relações e suas características de forma declarativa e direta, traduzindo-as automaticamente para os axiomas OWL correspondentes). Por fim, a persistência e o raciocínio lógico são suportados por um servidor de triplas RDF (exemplo: Apache Jena Fuseki), que opera como triplestore e motor de inferência, permitindo a execução de consultas SPARQL complexas sobre a base de conhecimento interconectada.

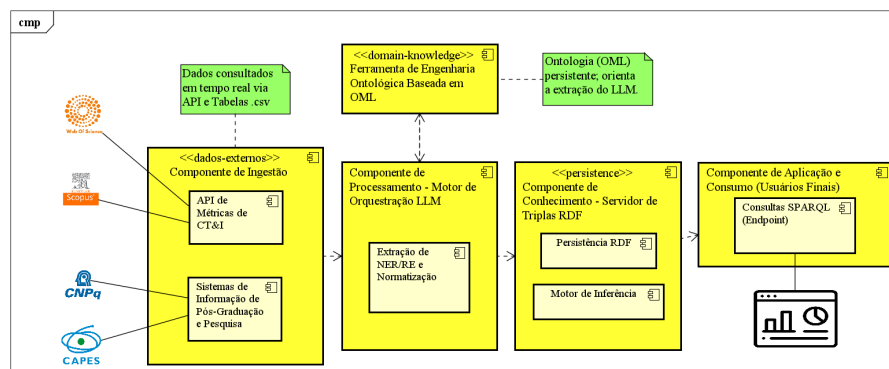


Figura 1. Visão Arquitetural de Componentes

O diagrama apresentado na Figura 2 sintetiza a conversão de dados heterogêneos (A) em inteligência acionável (B), permitindo visualizar o caminho desde a construção da consulta SPARQL (C) até a sua materialização em um dashboard analítico (D). Essa interface de usuário transforma triplas RDF abstratas em representações visuais interativas, provendo uma ferramenta de monitoramento que supera as limitações de relatórios estáticos ao permitir a exploração granular de indicadores de inovação e pesquisa.

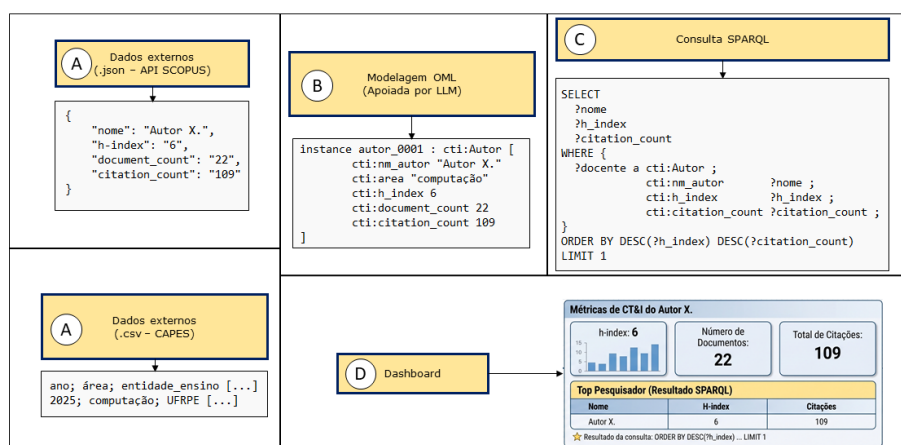


Figura 2. Exemplos de Dados Externos, Modelagem OML e Consultas SPARQL

3.2. Componente de Ingestão: Hibridismo e Vivacidade dos Dados

O Componente de Ingestão é responsável pela captura e padronização de fluxos de dados heterogêneos. A principal vantagem da abordagem híbrida é: enquanto dados de alto desempenho e métricas normalizadas, como *Field-Weighted Citation Impact* (FWCI) e contagem de citações, são consumidos via API de métricas de CT&I, os metadados governamentais e biográficos (oriundos de sistemas de informação de pós-graduação e pesquisa como: Plataforma Sucupira e Currículo Lattes) são integrados para fornecer o contexto institucional.

3.3. Componente de Processamento: O Pipeline Neuro-Simbólico

O núcleo da arquitetura reside no pipeline de processamento de linguagem natural mediado por LLMs. Nesta fase, o modelo transita do processamento probabilístico para a estruturação lógica. Utilizando técnicas de *few-shot prompting*, o LLM atua na Extração de Entidades Nomeadas (NER) e de Relações (RE), identificando conexões não explícitas, como o vínculo entre editais de fomento específicos e o surgimento de nichos tecnológicos emergentes.

Além da extração, este componente executa a normatização semântica (desambiguação), garantindo que variações de nomes de instituições ou tecnologias sejam mapeadas para identificadores únicos na ontologia de domínio. O LLM não apenas extrai o dado, mas gera uma justificativa textual sobre a relevância da conexão estabelecida, provendo a explicabilidade necessária para auditorias em políticas públicas.

3.4. Componente de Conhecimento: Persistência e Integridade Semântica

A informação processada é consolidada em uma estrutura de triplas Resource Description Framework (RDF), permitindo uma topologia de grafo conectada e propícia a algoritmos de inferência. Para mitigar o risco de alucinações inerente aos modelos generativos, este componente implementa a abordagem HITL.

A abordagem HITL neste componente visa garantir que, por exemplo, uma “Métrica de Impacto” esteja obrigatoriamente vinculada a uma “Publicação” ou “Pesquisador” existente no mundo real. Essa ancoragem assegura que a saída da GenAI seja

validada a fim de respeitar as regras de negócio do domínio de CT&I, transformando inferências estatísticas em conhecimento validado.

3.5. Componente de Aplicação e Consumo: Inteligência Estratégica

A interface de saída é desenhada para atender a dois perfis de tomada de decisão. Para analistas de dados, o sistema oferece um *SPARQL Endpoint*, permitindo consultas complexas que seriam inviáveis em bancos relacionais, como a identificação de pesquisadores com alto impacto ($FWCI > 1.2$) que ainda não possuem patentes registradas em áreas prioritárias.

Para o nível executivo, o modelo alimenta Dashboards Estratégicos baseados na topologia do grafo. Essas visualizações permitem identificar “ilhas de excelência” e lacunas de investimento de forma temática, facilitando a alocação de recursos baseada em evidências de conectividade e impacto científico.

4. Discussão e Limitações

A transição de um modelo de gestão de dados passivo para um ecossistema de conhecimento ativo, conforme proposto nesta arquitetura, representa uma mudança de paradigma na governança de CT&I. A discussão a seguir fundamenta-se na premissa de que a eficácia da decisão governamental não depende apenas da posse do dado, mas da capacidade de extrair dele uma semântica acionável que os modelos relacionais tradicionais falham em capturar [Bellomarini et al. 2022].

4.1. Discussão: O Valor Estratégico da Abordagem Neuro-Simbólica

Espera-se que a implementação do modelo conceitual proposto ajude a mitigar o problema da fragmentação informacional: as “ilhas de dados” mencionadas por [Arya et al. 2026]. Ao integrar métricas de CT&I com dados biográficos e de fomento via Grafos de Conhecimento (KGs), o modelo permite o que [Ouzounis et al. 2023] definem como data-driven ranking e visão holística. Diferente de uma consulta SQL em bancos tradicionais, que exigiria joins complexos e rígidos, a estrutura de KG permite navegar por relações implícitas. Por exemplo, a identificação de um nicho de pesquisa promissor não viria apenas da contagem de citações (FWCI), mas da conexão semântica entre editais de fomento específicos, a emergência de novas terminologias extraídas por LLMs e a rede de colaboração internacional dos pesquisadores envolvidos.

A principal contribuição teórica deste modelo reside na sua natureza neuro-simbólica. Como argumentam [Benjira et al. 2025, Wu et al. 2025], a combinação do poder indutivo dos LLMs com o rigor lógico das ontologias mitiga a opacidade dos sistemas “caixa-preta”. No contexto público, isso significa que o gestor pode auditar as sugestões do sistema, baseando-se na topologia do grafo e não apenas em probabilidades estatísticas [Malec et al. 2023].

O modelo não propõe uma migração massiva e custosa de dados, mas um componente semântico que interroga as fontes originais em tempo real. Para a administração pública, isso reduz o risco de decisões baseadas em dados obsoletos e evita a criação de novos data swamps, um problema crítico identificado por [Dreger et al. 2025].

A análise das métricas de CT&I a partir do modelo proposto permite que agências

de fomento e outras instituições governamentais identifiquem, com precisão, quais ecossistemas de inovação apresentam maior retorno sobre o investimento. Em um cenário de restrição orçamentária, a capacidade de eliminar redundâncias informacionais e mitigar a fragmentação decisória não é apenas uma melhoria operacional, mas um imperativo para a soberania tecnológica e a eficiência administrativa, garantindo que cada recurso público seja direcionado a áreas de real potencial transformador.

4.2. Limitações e Desafios de Implementação

Apesar da capacidade de transformação dos processos, a viabilidade do modelo enfrenta limitações que devem ser discutidas com rigor técnico. A primeira reside na dependência da qualidade e disponibilidade dos dados de origem. Como alertado por [Dessì et al. 2021], se os metadados da API de métricas de CT&I ou das plataformas governamentais apresentarem inconsistências ou lacunas de cobertura, o KG herdará essas falhas. A autoridade do grafo é diretamente proporcional à integridade das APIs que o alimentam.

Uma limitação intrínseca ao uso de GenAI é o risco de alucinações no povoamento do grafo, pois a automação total da extração de entidades (NER) e relações (RE) ainda é um desafio. A necessidade de uma abordagem HITL, conforme sugerido por [Doumanas et al. 2025], implica que especialistas de domínio ainda precisam validar as triplas de conhecimento mais críticas, o que pode se tornar um gargalo operacional em instituições públicas com quadros técnicos reduzidos.

Do ponto de vista computacional, a viabilidade econômica de modelos comerciais de LLM representam uma barreira de escalabilidade financeira. Manter um KG atualizado (dependendo de sua escala) via GenAI exige uma infraestrutura de custos que muitas vezes não está prevista nos orçamentos de TI governamentais.

Por fim, há a barreira da complexidade da modelagem. A construção de uma ontologia de domínio que seja, ao mesmo tempo, rigorosa o suficiente para garantir a inferência lógica e flexível o suficiente para acompanhar a evolução da ciência [Shimizu and Hitzler 2025] é uma tarefa de engenharia de software altamente especializada. A resistência organizacional à transição de sistemas legados para tecnologias de grafos, somada às implicações éticas sobre a transparência dos algoritmos de recomendação em políticas públicas, sugere que a implementação deste modelo exige não apenas maturidade tecnológica, mas também uma revisão dos marcos regulatórios de governança de dados no setor público.

5. Conclusão

A complexidade inerente à gestão de políticas públicas de CT&I exige uma transição paradigmática: o abandono de repositórios passivos e fragmentados em favor de ecossistemas de dados ativos e semanticamente interconectados. Este artigo cumpriu seu objetivo central ao propor e formalizar um modelo conceitual neuro-simbólico que utilize LLMs para a automação do povoamento de Grafos de Conhecimento, garantindo a integridade semântica dos dados de CT&I através de restrições ontológicas.

A principal contribuição deste estudo reside na arquitetura neuro-simbólica proposta. Ao combinar o poder indutivo e a flexibilidade dos LLMs com o rigor lógico e

a rastreabilidade das ontologias formais, o modelo oferece vantagens comparativas em relação às limitações dos bancos de dados relacionais tradicionais. Como discutido ao longo do trabalho, a incapacidade dos modelos tabulares em lidar com a heterogeneidade dos metadados científicos frequentemente resulta em data swamps, conforme alertado por [Dreger et al. 2025].

A proposta apresentada permite a realização de data-driven ranking e a identificação de nichos de pesquisa promissores que seriam invisíveis em análises puramente estatísticas. A inteligência estratégica governamental não deve depender apenas da posse do dado, mas da capacidade de realizar inferências complexas sobre ele [Ouzounis et al. 2023, Bellomarini et al. 2022]. O modelo proposto atende a essa demanda ao fornecer uma infraestrutura que suporta a descoberta de conhecimento e a simulação de cenários para a tomada de decisão baseada em evidências.

Apesar do potencial transformador, o modelo conceitual apresenta limitações que devem ser reconhecidas. A eficácia da arquitetura é intrinsecamente dependente da qualidade e da cobertura dos dados de origem. Inconsistências nos metadados da API de métricas de CT&I podem propagar erros para o grafo, reforçando a necessidade de mecanismos de curadoria e a abordagem HITL para validar triplas de conhecimento, conforme sugerido por [Doumanas et al. 2025]. Além disso, a implementação em larga escala enfrenta desafios operacionais, incluindo o custo computacional de tokens de LLMs e a necessidade de adaptação da ontologia de domínio aos contextos específicos de diferentes agências de fomento ou realidades regionais.

Como trabalhos futuros, planeja-se a implementação de um protótipo funcional aprimorando a pilha tecnológica descrita. Este protótipo será submetido a ciclos de validação com gestores públicos reais, visando avaliar a usabilidade dos dashboards estratégicos e a precisão das consultas SPARQL em cenários de auditoria e planejamento de editais. A evolução deste modelo deve contemplar, ainda, a integração de técnicas de RAG e linguagens de restrição como SHACL para reduzir alucinações e garantir que as respostas do sistema estejam sempre ancoradas na base de conhecimento verificada.

Dessa forma, a unificação de Grafos de Conhecimento e GenAI representa um avanço para a soberania estratégica nacional em CT&I. Ao prover transparência lógica e capacidade de inferência, o modelo proposto não apenas moderniza a infraestrutura tecnológica governamental, mas fortalece a responsabilidade na alocação de recursos públicos, garantindo que o ciclo de políticas de inovação seja guiado por uma compreensão profunda e conectada do ecossistema científico.

6. Declaração sobre Uso de IA Generativa

A concepção teórica, o delineamento metodológico e a arquitetura original proposta neste trabalho foram desenvolvidos integralmente pelos autores. O modelo de linguagem Gemini 3 Flash (*gemini-3-flash-preview*) foi utilizado exclusivamente como ferramenta de suporte para o refinamento da redação técnica e a otimização da clareza descritiva dos componentes. O uso da IA limitou-se ao auxílio na articulação lógica e na precisão estilística da arquitetura predefinida pelos autores, que assumem integral responsabilidade pela originalidade, rigor científico e conteúdo intelectual do estudo.

Referências

- Abu-Salih, B., Alotaibi, S., Alanazi, A. L., Abu Khurma, R., Al-Shboul, B., Khouri, A., and Aljaafari, M. (2025). Using large language models for semantic interoperability: A systematic literature review. *ICT Express*, 11(4):819–837. <https://doi.org/10.1016/j.icte.2025.06.011>.
- Arya, V., Saraf, A., Chichkanov, N., Papa, A., and Romano, M. (2026). AI-enhanced competency transfer hubs: a conceptual framework for university-industry engagement and knowledge sharing. *The Journal of Technology Transfer*, 51(2):682–712. <https://doi.org/10.1007/s10961-025-10233-7>.
- Bellomarini, L., Fayzrakhmanov, R. R., Gottlob, G., Kravchenko, A., Laurenza, E., Nenov, Y., Reissfelder, S., Sallinger, E., Sherkhonov, E., Vahdati, S., and Wu, L. (2022). Data science with Vadalog: Knowledge Graphs with machine learning and reasoning in practice. *Future Generation Computer Systems*, 129:407–422. <https://doi.org/10.1016/j.future.2021.10.021>.
- Benjira, W., Atigui, F., Bucher, B., Grim-Yefsah, M., and Travers, N. (2025). Automated mapping between SDG indicators and open data: An LLM-augmented knowledge graph approach. *Data & Knowledge Engineering*, 156:102405. <https://doi.org/10.1016/j.datak.2024.102405>.
- Bergek, A., Hellsmark, H., and Karltorp, K. (2023). Directionality challenges for transformative innovation policy: lessons from implementing climate goals in the process industry. *Industry and Innovation*, 30(8):1110–1139. <https://doi.org/10.1080/13662716.2022.2163882>.
- Colombo, A., Bernasconi, A., and Ceri, S. (2025). An LLM-assisted ETL pipeline to build a high-quality knowledge graph of the Italian legislation. *Information Processing & Management*, 62(4):104082. <https://doi.org/10.1016/j.ipm.2025.104082>.
- Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., and Motta, E. (2021). Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. *Future Generation Computer Systems*, 116:253–264. <https://doi.org/10.1016/j.future.2020.10.026>.
- Doumanas, D., Ntalouka, E., Vassilakis, C., Wallace, M., and Kotis, K. (2025). Stitching History into Semantics: LLM-Supported Knowledge Graph Engineering for 19th-Century Greek Bookbinding. *Machine Learning and Knowledge Extraction*, 7(3):59. <https://doi.org/10.3390/make7030059>.
- Dreger, M., Malek, K., and Eikerling, M. (2025). Large language models for knowledge graph extraction from tables in materials science. *Digital Discovery*, 4(5):1221–1231. <https://doi.org/10.1039/D4DD00362D>.
- Gokhberg, L., Fursov, K., Miles, I., and Perani, G. (2013). Chapter 15: Developing and using indicators of emerging and enabling technologies. In *Handbook of Innovation Indicators and Measurement*. Edward Elgar Publishing. Section: Handbook of Innovation Indicators and Measurement. <https://doi.org/10.4337/9780857933652.00027>.
- Hoseini, S., Theissen-Lipp, J., and Quix, C. (2024). A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. *Journal of Web Semantics*, 81:100819. <https://doi.org/10.1016/j.websem.2024.100819>.

- Kalaycı, T. E., Bricelj, B., Lah, M., Pichler, F., Scharrer, M. K., and Rubeša-Zrim, J. (2021). A Knowledge Graph-Based Data Integration Framework Applied to Battery Data Management. *Sustainability*, 13(3):1583. <https://doi.org/10.3390/su13031583>.
- Malec, S. A., Taneja, S. B., Albert, S. M., Elizabeth Shaaban, C., Karim, H. T., Levine, A. S., Munro, P., Callahan, T. J., and Boyce, R. D. (2023). Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: A use case studying depression as a risk factor for Alzheimer’s disease. *Journal of Biomedical Informatics*, 142:104368. <https://doi.org/10.1016/j.jbi.2023.104368>.
- Meroño-Peñuela, A., Simperl, E., Kurteva, A., and Reklos, I. (2025). KG.GOV: Knowledge graphs as the backbone of data governance in AI. *Journal of Web Semantics*, 85:100847. <https://doi.org/10.1016/j.websem.2024.100847>.
- Ouzounis, S., Kanterakis, A., Panagiotopoulos, V., Cavouras, D., Zoumpoulakis, P., Matsoukas, M.-T., Katsila, T., and Kalatzis, I. (2023). Data-Driven Drug Repurposing in Diabetes Mellitus through an Enhanced Knowledge Graph. *Engineering Proceedings*, 50(1):9. <https://doi.org/10.3390/engproc2023050009>.
- Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K. A., Good, B. M., Balhoff, J. P., Fontana, T., Blau, H., Matentzoglou, N., Harris, N. L., Munoz-Torres, M. C., Haendel, M. A., Robinson, P. N., Joachimiak, M. P., and Mungall, C. J. (2021). KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns*, 2(1). <https://doi.org/10.1016/j.patter.2020.100155>.
- Shimizu, C. and Hitzler, P. (2025). Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, 85:100862. <https://doi.org/10.1016/j.websem.2025.100862>.
- Vidal, M.-E., Chudasama, Y., Huang, H., Purohit, D., and Torrente, M. (2025). Integrating Knowledge Graphs with Symbolic AI: The Path to Interpretable Hybrid AI Systems in Medicine. *Journal of Web Semantics*, 84:100856. <https://doi.org/10.1016/j.websem.2024.100856>.
- Wu, L.-x., Jiang, Y., Luo, T.-y., Hou, J.-x., Deng, Y., Han, L.-x., Jiang, T.-f., and Bao, J. (2025). Interpretable AI-assisted diagnosis of papillary thyroid cancer cytopathology using graph neural networks and knowledge graphs. *Scientific Reports*, 15(1):32165. <https://doi.org/10.1038/s41598-025-18235-z>.