AKR Advances on Knowledge Representation Journal

Dados sintéticos para treinamento de Inteligência Artificial: intersecções com a Organização e Representação da Informação e do Conhecimento

Ananda Fernanda de Jesus[1], Wesley Algarve[2], José Eduardo Santarem Segundo[3]

Resumo. *Enquanto os dados reais são coletados de diferentes contextos do mundo real — como a observação de fenômenos diversos ou a interação de pessoas com sistemas e máquinas — os dados sintéticos são gerados por meio da aplicação de diferentes técnicas, com o objetivo de ampliar, equilibrar ou diversificar um determinado conjunto de dados. O uso de dados sintéticos não é recente: há muito tempo é empregado em análises estatísticas e, com frequência, no campo da Inteligência Artificial. No entanto, o atual contexto tecnológico, marcado pela proliferação de Grandes Modelos de Linguagem e pela popularização das Inteligências Artificiais Generativas, tem evidenciado a necessidade de volumes massivos de dados para o treinamento e a evolução contínua desses modelos, o que tem ampliado significativamente a adoção de dados sintéticos no treinamento de diferentes tipos de Inteligência Artificial. Reconhecendo a importância dos dados sintéticos nesse cenário, a presente pesquisa busca apresentar e discutir conceitos relacionados a esse tipo de dado, bem como identificar suas possíveis interfaces com as áreas de Organização e Representação da Informação e do Conhecimento. Para isso, parte-se da identificação e discussão dos conceitos pertinentes, por meio de um estudo de terminologia pontual e de uma análise exploratória baseada em um protocolo de pesquisa. Como resultado, espera-se traçar relações entre o contexto dos dados sintéticos e a representação da informação e a organização do conhecimento, identificando tanto os desafios que esses dados impõem à representação quanto*

[1] Doutora em Ciência da Informação pela Universidade Estadual Paulista (UNESP). Mestre em Ciência da Informação pela Universidade Federal de São Carlos (UFSCar). Bolsa estágio de pesquisa no exterior (BEPE) na Universidad de La República (UDELAR). Bacharel em Biblioteconomia e Ciência da Informação pela UFSCar.
E-mail: af.jesus@unesp.br. ORCID: https://orcid.org/0000-0001-7873-6040

[2] Doutorando e Mestre em Ciência da Informação pela Universidade Estadual Paulista (UNESP). Bacharel em Biblioteconomia e Ciência da Informação pela Universidade Federal de São Carlos (UFSCar).
E-mail: wesley.algarve@unesp.br. ORCID: https://orcid.org/0000-0003-3528-8510

[3] Livre Docente em Informação e Tecnologia pela Universidade de São Paulo (USP), 2020. Pós-Doutorado pela Faculdade de Engenharia da Computação da Western University/Canadá, 2018. Doutor e Mestre em Ciência da Informação pela Universidade Estadual Paulista (UNESP); Professor Doutor no Departamento de Educação, Informação e Comunicação, da USP; Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/Marília na linha de Informação e Tecnologia.
E-mail: santarem@usp.br. ORCID: https://orcid.org/0000-0003-3360-7872

as potenciais contribuições que a representação e a organização podem oferecer a esse universo.

Palavras-chave: *Dados sintéticos; Representação da informação; Organização do conhecimento; Inteligência Artificial; Grandes modelos de linguagem.*

1. Introdução

O treinamento de sistemas baseados em Inteligência Artificial (IA) demandam grandes volumes de dados, cuja obtenção constitui-se um dos principais desafios enfrentados pelos desenvolvedores desses sistemas.

O avanço na produção sistemas e ferramentas de IA se deve em partes à melhores capacidades de processamento de grandes quantidades de dados, que são utilizados em seu treinamento. Dwivedi (2023) destaca que no estágio atual desse tipo de ferramenta, a geração, ou criação, ainda é limitada a uma recombinação dos conteúdos dos dados do treinamento, tendo como base análise probabilística.

Grandes conjuntos de dados são complexos, provenientes da agregação de fontes heterogêneas de dados, possuem diferentes formatos e estruturas, estão sujeitos a erros, inconsistências e ruídos (Fan *et al.*, 2014).

Quando classificados e categorizados, os dados podem ser desbalanceados:

> Um conjunto de dados é dito desbalanceado quando nele existe uma clara desproporção entre número de exemplos de uma ou mais classe em relação as demais classes. Pode-se observar, por exemplo, em um estudo de caso de uma determinada doença rara em uma população, que o número de pessoas portadoras da doença é muito menor do que o número de não portadores, ou seja, existe uma grande desproporção entre o número de exemplos das classes. (Barella, 2015, p. 2-3)

Esse desbalanceamento está presente na maioria dos conjuntos de dados reais e, quando aplicados no treinamento da IA, podem gerar uma série de problemáticas, como a tendência ao favorecimento de classes majoritárias, imprecisão dos resultados ou a perpetuação de discriminação por meio do apagamento das informações relativas a classes minoritárias.

A aplicação de dados desbalanceados ou que possuem algum tipo de viés intrínseco podem não só perpetuar, como amplificar questões discriminatórias, refletindo preconceitos e desigualdades sociais, como gênero, raça, orientação sexual e classe social (Coneglian *et al.*, 2017).

Outro desafio em relação a aplicação de dados no treinamento de IAs está relacionado à privacidade, especialmente no contexto de dados pessoais e dados

sensíveis. “Em uma era de crescente coleta e análise de dados, estabelecer medidas diferenciais de privacidade é essencial para manter a confiança e a transparência, bem como proteger os direitos individuais de privacidade” (Goyal & Mahmoud, 2024, p. 2).

Destaca-se ainda as previsões a respeito de um potencial esgotamento de fontes de dados reais disponíveis para o treinamento de IAs:

> Estimamos o estoque de texto público gerado por humanos em cerca de 300 trilhões de tokens. Se as tendências continuarem, os modelos de linguagem utilizarão totalmente esse estoque entre 2026 e 2032, ou até antes, se intensamente sobrecarregados. (Villalobos *et al.*, 2024, não paginado, tradução nossa)

Em síntese, a criação de um *corpus* para treinamento de IAs é perpassada por uma série de desafios relacionados a dados desbalanceados, previsão de esgotamento de novos dados em quantidade suficiente para avanços, aspectos relacionados à privacidade dos dados, a violação de direitos autorais, e a suscetibilidade à criação de *corpus* de dados enviesados que amplificam aspectos discriminatórios.

Uma das possibilidades para lidar com esses desafios é a geração e aplicação de dados sintéticos.

> A geração de dados sintéticos surgiu como uma solução promissora para superar os desafios impostos pela escassez de dados e preocupações com privacidade, bem como para atender à necessidade de treinar algoritmos de inteligência artificial (IA) em dados imparciais com tamanho de amostra e poder estatístico suficientes. (Pezoulas *et al.*, 2024, p. 2892, tradução nossa)

Compreendendo a importância dos dados sintéticos no contexto vigente, e buscando contribuições para o seu desenvolvimento relacionadas à Ciência da Informação, a presente pesquisa parte do questionamento: Qual a relação entre dados sintéticos criados para treinamento de IA e as áreas da Organização e Representação e da Informação e do Conhecimento?

A pesquisa teve como objetivo discutir os principais conceitos relacionados à criação e aplicação de dados sintéticos e abordar seu potencial relação com a Organização e Representação e da Informação e do Conhecimento. Para atingir esse objetivo, foi conduzida uma pesquisa dividida em duas etapas principais, baseada em análise exploratória e adoção de protocolo de pesquisa. A próxima seção apresenta os procedimentos metodológicos adotados.

2. Procedimentos metodológicos

A presente pesquisa se caracteriza como exploratória e descritiva, considerando-se que busca explorar e discutir o potencial de colaboração mútua entre a Representação da Informação/ Organização do Conhecimento e Ciência da Computação, no processo de criação e aplicação de dados sintéticos no treinamento de ferramentas de IA. Em relação a seus resultados, trata-se de uma pesquisa teórica e qualitativa.

A condução da pesquisa foi dividida em duas etapas principais: 1) Análise dos aspectos teóricos e processuais relacionados a dados sintéticos no contexto da IA; 2) Análise exploratória a respeito da relação entre dados sintéticos e a Representação da Informação e Organização do Conhecimento.

Nessa etapa, buscou-se por referencial consolidado a respeito da criação, aplicação, representação e recuperação de dados sintéticos no contexto de sua aplicação em ferramentas de IA.

Nessa etapa, considerou as palavras-chave: Dados Sintéticos; ciclo de vida de dados sintéticos; criação de dados sintéticos; representação de dados sintéticos; reuso de dados sintéticos; dados sintéticos para inteligência artificial; dados sintéticos para inteligência artificial generativa. Foram consideradas as bases *Web of Science*; LISTA; *Semantic Scholar* e Google Acadêmico.

Com isso, buscou-se identificar:

- Definição do termo “dados sintéticos”;
- Necessidade e relevância de dados sintéticos no contexto da IA;
- Principais formas de criação de dados sintéticos;
- Ciclo de vida dos dados sintéticos.

Construído esse referencial, partiu-se para a análise da relação entre Representação da Informação e Organização do Conhecimento e os dados sintéticos, baseada na condução de análise exploratória da literatura baseada em protocolo de pesquisa. Ressalta-se que esse tipo de abordagem não tem como objetivo exaurir de maneira metódica a literatura, mas sim fornecer um panorama amplo de como o assunto vem sendo explorado no escopo selecionado, bem como aportes para iniciar as discussões a respeito da temática.

Para orientar a condução do levantamento, foi estabelecido o protocolo apresentado no quadro 1.

Quadro 1 – Sistematização dos conceitos de OI, RI, OC e RC

Protocolo de busca	
Título	Dados sintéticos para treinamento de Inteligência Artificial: intersecções com a representação da informação e organização do conhecimento
Objetivos	Identificar a potencial relação entre dados sintéticos e a Representação da Informação e Organização do Conhecimento
Pergunta de pesquisa (principal)	Como a relação entre dados sintéticos e a representação da informação e do conhecimento tem sido abordada pela literatura?
Estratégia de busca	"synthetic data" AND ("representation of information" OR "knowledge representation" OR "knowledge organization" OR "information organization")
Bases de dados consultada	Web of Science; LISTA; Semantic Scholar e Google Acadêmico
Idiomas	Português; Inglês e Espanhol
Critérios de Inclusão	(I) Aborda a relação entre dados sintéticos e a representação da informação e do conhecimento (I) Foco na relação de dados sintéticos e a representação da informação e do conhecimento (I) Aplicações de dados sintéticos na representação da informação e do conhecimento (I) Aplicações de instrumentos da representação da informação e do conhecimento no contexto de dados sintéticos
Critérios de exclusão	(E) Não está nos idiomas estabelecidos para a pesquisa (E) Não aborda a temática de interesse (E) Não foi possível obter acesso ao documento completo
Formulário de extração	1. Definição de dados sintéticos 2. Relação entre dados sintéticos e a representação da informação e do conhecimento 3. Desafios relacionados ao contexto 4. Aplicações

Fonte: Autores (2026)

Apresentados os procedimentos adotados, a próxima seção apresenta os resultados obtidos.

3. Organização e Representação da Informação e do Conhecimento: instrumentos, práticas e procedimentos

Essa seção foi construída com o objetivo de abordar os principais instrumentos, práticas e procedimentos relacionados à Organização e Representação da Informação e do Conhecimento, com o propósito de promover a discussão de suas possíveis interações

com o contexto da produção, representação, recuperação e aplicação de dados sintéticos no treinamento de IA.

Para isso, torna-se necessária a discussão dos termos Organização da Informação (OI), Representação da Informação (RI), Organização do Conhecimento (OC) e Representação do Conhecimento (RC).

Existe, na literatura da Ciência da Informação, uma pluralidade terminológica em relação ao emprego desses termos, como aponta Lima (2020, p. 62): "O termo Organização e Representação do Conhecimento e da Informação possui uma dispersão terminológica, sendo que algumas vezes são utilizados como complementares, em outras são utilizados de maneira distinta".

Embora não seja o objetivo do presente trabalho discutir exaustivamente os limites que diferenciam as áreas mencionadas e as fronteiras entre suas definições conceituais, buscou-se discuti-los, visando identificar seus objetivos, principais processos e os instrumentos a elas relacionados.

Tomando como base as discussões realizadas por Brascher & Café (2008) e Lima (2020), a figura 1 apresenta uma sistematização dos conceitos de OI, RI, OC e RC.

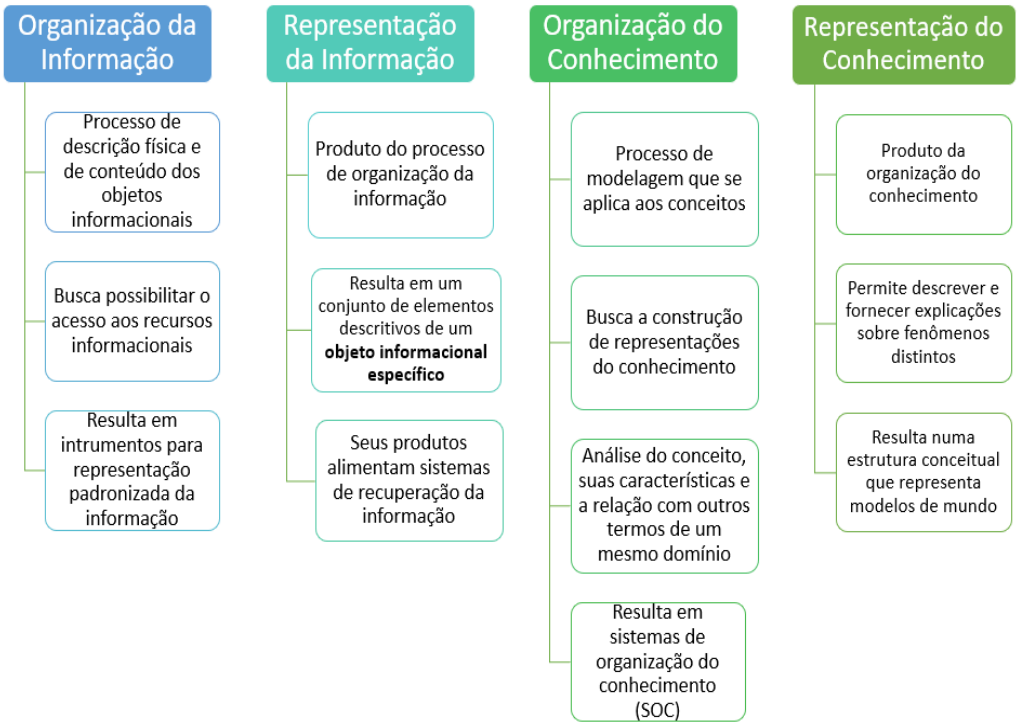


Figura 1 – Sistematização dos conceitos de OI, RI, OC e RC
Fonte: Autores (2026), baseado em Brascher & Café (2008) e Lima (2020).

Das interrelações dessas áreas são conduzidos processos como a representação descritiva e temática dos recursos informacionais, a criação de resumos, a modelagem de domínios, e os estudos terminológicos.

Esses processos permitem a organização e representação dos conteúdos dos recursos e seu agrupamento em classes, por meio da organização e representação do conhecimento. Permitem ainda a representação individual dos recursos informacionais pertencentes às diferentes classes, possibilitando a criação de sistemas de recuperação da informação, como catálogos, buscadores e bases de dados.

No âmbito da Organização e Representação do Conhecimento, destacam-se os Sistemas de Organização do Conhecimento (SOC).

> A representação do conhecimento é feita por meio de diferentes tipos de sistemas de organização do conhecimento (SOC) que são sistemas conceituais que representam determinado domínio por meio da sistematização dos conceitos e das relações semânticas que se estabelecem entre eles. (Brascher & Café, 2008, p. 8)

Os SOCs se estabeleceram em um contexto de explosão da informação, primeiro para lidar com recursos basicamente bibliográficos, e depois para contextos científicos, como periódicos e publicações seriadas (Marcondes, 2021).

SOCs variam em relação a sua estrutura, nomenclatura e complexidade. A figura 2 apresenta uma sistematização dos principais tipos de SOC:

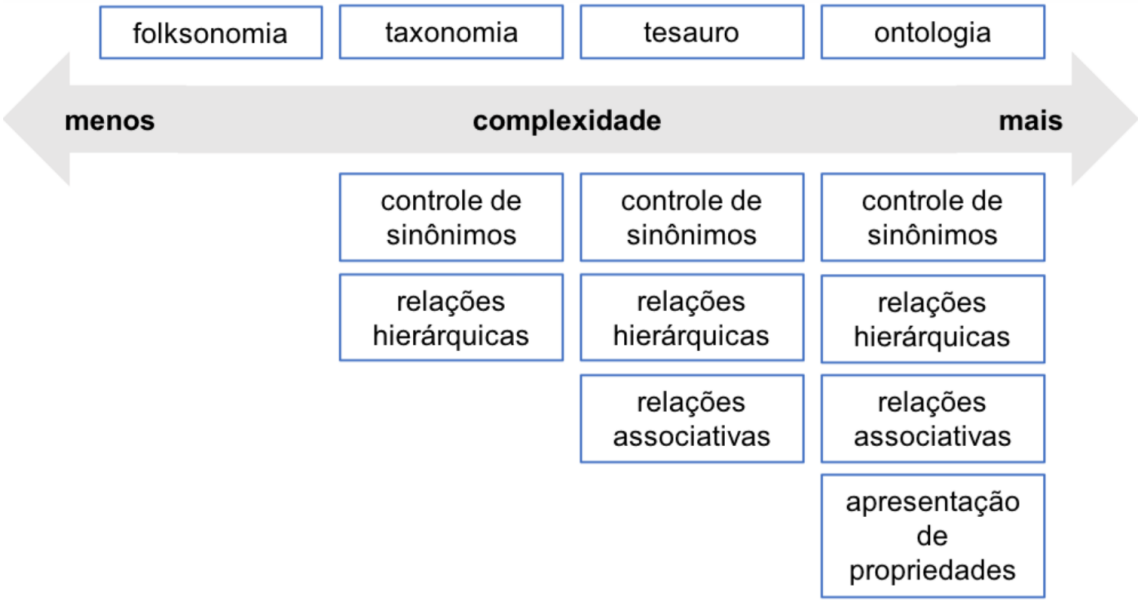


Figura 2 – Complexidade dos SOC
Fonte: (Moreira, 2018, p. 102).

No âmbito da Organização e Representação da Informação descritiva:

> Os elementos que compõem a representação descritiva são: (a) a estrutura dos campos; (b) a descrição do item; (c) os pontos de acesso descritivos. Esses elementos constituem a ficha catalográfica com a descrição física do item, indicando as relações bibliográficas com elementos descritivos. Na representação da informação, os metadados atuam como referenciais ao item informacional representado, e como intermediários entre o objeto representado e o usuário. (Lima, 2020, p. 80)

Os metadados e padrões de metadados possuem muita relevância na representação e recuperação da informação, dentro e fora do universo bibliográfico, e passaram a ter maior destaque por seu papel na recuperação da informação em ambientes digitais.

> Metadados são atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação. Os padrões de metadados são estruturas de descrição constituídas por um conjunto predeterminado de metadados (atributos codificados ou identificadores de uma entidade) metodologicamente construídos e padronizados. (Alves, 2010, p. 48)

Quando se considera o contexto *web*, a barreira entre função de representação da informação ou do conhecimento dos instrumentos e processos se torna ainda mais difusa. Um exemplo dessa complexidade são os vocabulários semânticos e as ontologias.

Os Vocabulários Semânticos são estruturados com base nas melhores práticas do *Linked Data*, sendo elaborados visando o mapeamento das propriedades e da relação entre as entidades de determinado domínio, permitindo a explicitação dessas relações de maneira formal (legível por máquina), mas a depender de sua aplicação, também fornecem a estrutura necessária para a representação descritiva da informação.

Além do contexto digital, soma-se a essa complexidade o objeto de representação e organização em questão: os dados. Um dado pode ser entendido como “a unidade de conteúdo de granularidade mais fina possível de determinado contexto de uso" (Santos & Sant’Ana, 2013, p. 205).

Os dados possuem características próprias que influenciam no seu processo de organização e representação, são representados em diferentes ciclos de vida, e a sua compreensão depende da sua associação a um modelo que o contextualiza, como aponta Jesus (2025, p. 98):

> Em seu aspecto abstrato, o dado se refere a entidades do mundo real (conceituais ou físicas) e a relação existente entre elas, podendo um dado possuir mais de uma representação. Para

> permitir a sua interpretação, o dado depende de um modelo, abstração do domínio no qual se insere, podendo esse modelo estar implícito ou explícito. O modelo estabelece as entidades (e) do domínio, os atributos (a) dessa entidade e os valores (v) possíveis/esperados desse atributo para a entidade em questão. Nesse sentido, cada dado equivale a uma tripla. Quando registrado, seja em meio analógico ou digital, o dado se torna passível de armazenamento, recuperação, processamento (por humanos ou máquinas), interpretação e reinterpretação e avaliação de qualidade.

Dados sintéticos não são resultado dos processos de mensuração ou observação, assim como não representam diretamente objetos, conceitos e eventos do mundo real. Eles são produzidos com um propósito pré-estabelecido, buscando simular esses fenômenos, por meio da identificação de padrões nos dados que atuam como base para a sua criação ou no mapeamento de padrões de comportamento.

Nesse sentido, para que possam ser adequadamente representados e organizados, torna-se necessário entender os aspectos relacionados aos processos utilizados na sua produção e aplicação, os contextos da produção e os dados originais utilizados como base, quando esse fator se mostrar pertinente.

4. Dados sintéticos

Os dados sintéticos podem ser definidos como dados produzidos artificialmente de forma a serem utilizados no treinamento de sistemas em contextos nos quais a obtenção de dados reais é inviável, limitada ou onerosa (Nikolenko, 2021). Nesse sentido, a "produção artificial" refere-se a dados que não foram coletados nem extraídos de sujeitos ou eventos do mundo real, mas sim gerados por meio de sistemas algorítmicos capazes de reproduzir determinadas características dos dados reais (Jordon *et al.*, 2022). Em outras palavras, os dados sintéticos são criados e utilizados para mimetizar e substituir os dados do mundo real, de forma a serem empregados para treinamento de algoritmos (Jacobsen, 2023).

A compreensão dessa definição exige, contudo, distinguir os dados sintéticos dos dados reais, especialmente no que se refere à sua origem, aos processos de produção e suas finalidades de uso. Enquanto dados reais são coletados de diferentes contextos do mundo real, derivados por exemplo, da observação de fenômenos distintos ou da interação de pessoas com diferentes sistemas e máquinas, os dados sintéticos são gerados inicialmente com base em dados reais, sendo aplicadas técnicas distintas em sua criação, visando ampliar, equilibrar e/ou diversificar o *corpus* a ser utilizados. "Dados sintéticos ajudam a aumentar dados escassos, minimizar vieses, preservar a privacidade de dados e simular cenários futuros" (Gartner, 2024, não paginado, tradução nossa).

A ascensão dos dados sintéticos tem sido impulsionada pela escassez de dados disponíveis para o treinamento de sistemas, principalmente os baseados em IA, decorrente de fatores como restrições associadas à privacidade, presença de informações sensíveis, insuficiência de dados rotulados e limitações na qualidade dos dados. Soma-se a esse cenário desafios relacionados a dados desbalanceados, a previsão de esgotamento de novos dados em volume suficiente para sustentar avanços significativos, bem como a suscetibilidade à criação de conjuntos de dados enviesados, capazes de amplificar padrões discriminatórios.

Atualmente, existem uma série de métodos avançados para a criação de dados sintéticos:

> Dados sintéticos são gerados programaticamente aplicando regras predeterminadas, construções lógicas e modelos de simulação para criar dados que se assemelham muito às qualidades e comportamentos de dados do mundo real. Esses métodos são altamente adaptáveis e podem ser ajustados para recriar os padrões, relacionamentos e estruturas específicos observados em dados reais. (Goyal & Mahmoud, 2024, p. 12, tradução nossa)

Uma das opções de geração é por meio da aplicação das técnicas de IAGen. Um estudo realizado pela Gartner (2024, não paginado), prevê que "até 2026, 75% das empresas utilizarão IA generativa para criar dados sintéticos de clientes, em comparação a menos de 5% em 2023".

Ou seja, nesse contexto, as Inteligências Artificiais Generativas (IAGen) atuam em uma espécie de ciclo onde podem ser empregadas para facilitar e otimizar as técnicas de geração de dados sintéticos, que por sua vez podem ser empregadas novamente em seu treinamento, bem como alimentar outras técnicas de IA.

Sob essa perspectiva, os benefícios da geração de dados sintéticos evidenciam-se tanto no plano da proteção de informações quanto na ampliação das possibilidades analíticas e operacionais.

> Dados sintéticos permitem que as empresas produzam conjuntos de dados realistas, mas fictícios, que retêm a estrutura e os padrões fundamentais reais, enquanto não revelam nenhuma informação sensível. Isso permite que as partes interessadas realizem análises, criem algoritmos e testem aplicativos, mantendo a privacidade e a confidencialidade individuais. Outro problema que os dados sintéticos abordam é o aumento de dados. Quando acadêmicos e empresas tentam desenvolver um aplicativo para um assunto extremamente específico, eles frequentemente acham difícil reunir um conjunto de dados completo. Técnicas de geração de dados sintéticos podem gerar novas instâncias de dados com atributos ou circunstâncias exclusivas que não são vistas no conjunto de dados original. (Goyal & Mahmoud, 2024, p. 2, tradução nossa)

A aplicação de dados sintéticos também facilita o desenvolvimento de pesquisas, ampliando as possibilidades de transparência, reuso e compartilhamento dos dados, especialmente em casos onde são utilizados dados sensíveis ou pessoais.

O uso de dados sintéticos é uma solução já adotada em abordagens estatísticas e utilizado a décadas no âmbito da IA, como, por exemplo, para lidar com conjuntos de dados desbalanceados no treinamento de modelos de *Machine Learning* (ML), utilizando técnicas como *oversampling*, que consiste em:

> Aumentar o número de instâncias de classes minoritárias para equilibrar a distribuição de classes. O oversampling mais simples é a sobre amostragem aleatória, que simplesmente duplica instâncias minoritárias. A fraqueza mais grave aqui é que ela não adiciona nenhuma informação nova ao conjunto de dados e pode causar sobre ajuste de classificadores. (Zheng *et al.*, 2015, p. 1019, tradução nossa)

Observa-se um crescimento significativo na incorporação de dados sintéticos ao treinamento de algoritmos em uma ampla gama de aplicações. Na literatura, seu uso tem sido reportado em contextos diversos, incluindo aplicações na área da saúde (Dahmen & Cook, 2019), sistemas de transmissão e análise esportiva automatizada (Chen & Little, 2019), análise automática de imagens de satélite e sensoriamento remoto (Shermeyer *et al.*, 2021), sistemas de reconhecimento facial (Boutros *et al.*, 2023) e identificação automatizada de espécies de peixes (Allken *et al.*, 2018). Esses exemplos evidenciam a transversalidade dos dados sintéticos como estratégia para suprir limitações, já mencionadas, associadas à disponibilidade, rotulagem e uso ético de dados reais em diferentes domínios de aplicação.

Embora possam solucionar uma série de limitações relacionadas com o uso de dados reais, as diferentes técnicas de geração de dados sintéticos estão sujeitas a limitações que podem afetar diretamente o funcionamento de sistemas de IA que os utilizem em seu treinamento. Goyal e Mahmoud (2024) citam como exemplos dessas limitações: falha em capturar a diversidade dos dados originais, tendência a priorizar a geração de dados que se assemelham muito aos originais, não acrescentando novas informações, necessidade de grande processamento e memória, com altos custos de implementação envolvidos, e exigência de grande volume de dados para um treinamento eficaz. Os autores destacam ainda que:

> Se os dados de treinamento forem tendenciosos, os dados sintéticos provavelmente refletirão esses vieses, potencialmente levando a resultados injustos ou discriminatórios quando usados em aplicações do mundo real. Além disso, garantir a privacidade e a segurança dos dados sintéticos é crucial, pois o manuseio inadequado pode levar ao vazamento de informações confidenciais. (Goyal & Mahmoud, 2024, p. 29, tradução nossa)

Outro aspecto a ser levado em consideração é o limitado poder de autoalimentação desses dados, que leva ao questionamento sobre quão inovadores podem ser sistemas constantemente alimentados com dados sintéticos, e se esses dados seguirão reproduzindo os padrões identificados nos dados reais.

Portanto, os dados sintéticos não devem ser considerados um substituto direto dos dados reais. Dados sintéticos são, por definição, versões distorcidas dos dados reais. Dessa forma, processos de modelagem ou inferência baseados exclusivamente em dados sintéticos estão sujeitos a riscos adicionais. Em alguns domínios, seu uso pode ser adequado e não gerar impactos significativos, em outros, entretanto, pode resultar em consequências graves. Assim, os resultados produzidos por sistemas que utilizam dados sintéticos devem ser analisados e interpretados de forma crítica e, sempre que necessário, validados e ajustados com base em dados reais (Jordon *et al.*, 2022; Håkansson & Phillips-Wren, 2024).

Apresentados a síntese das temáticas abordadas nessa pesquisa, a próxima seção detalha os resultados obtidos.

5. Resultados do levantamento

As buscas nas bases de dados resultaram em 69 artigos. Durante a etapa de identificação, constatou-se 12 artigos duplicados, sendo aceitos assim para a próxima etapa 57 artigos. Na segunda etapa, de elegibilidade, a introdução, metodologia e resultados foram lidos, aplicando-se tanto os critérios de inclusão quanto os de exclusão, sendo recusados 32 artigos e aceitos, ao final, 25 artigos que compuseram o corpus desta pesquisa.

A figura 3 mostra em detalhes a quantidade de artigos aceitos e excluídos em cada uma das etapas:

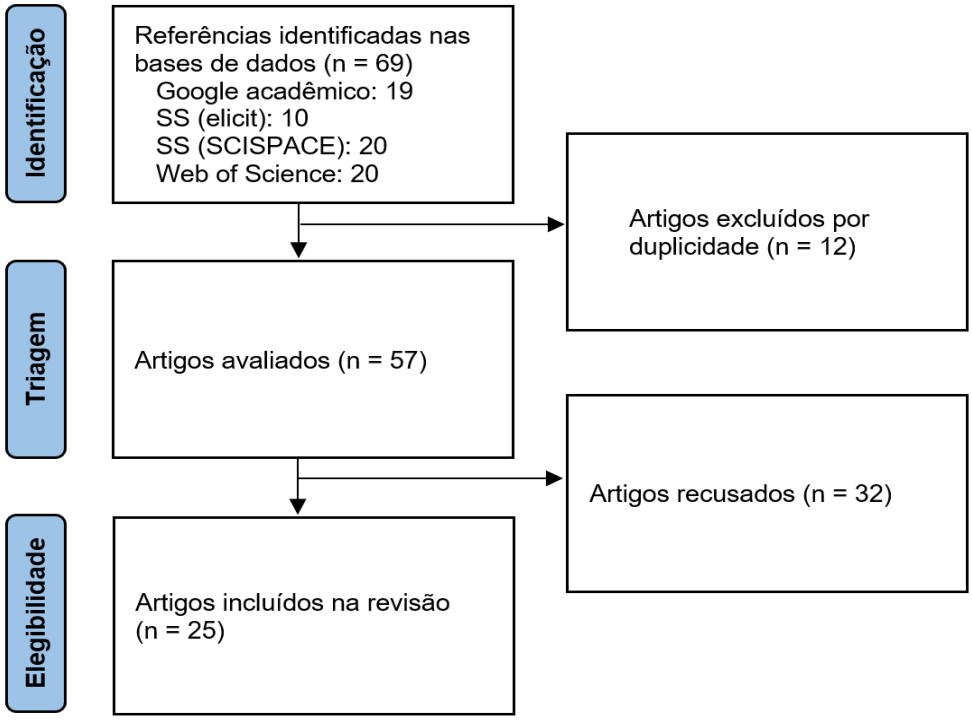
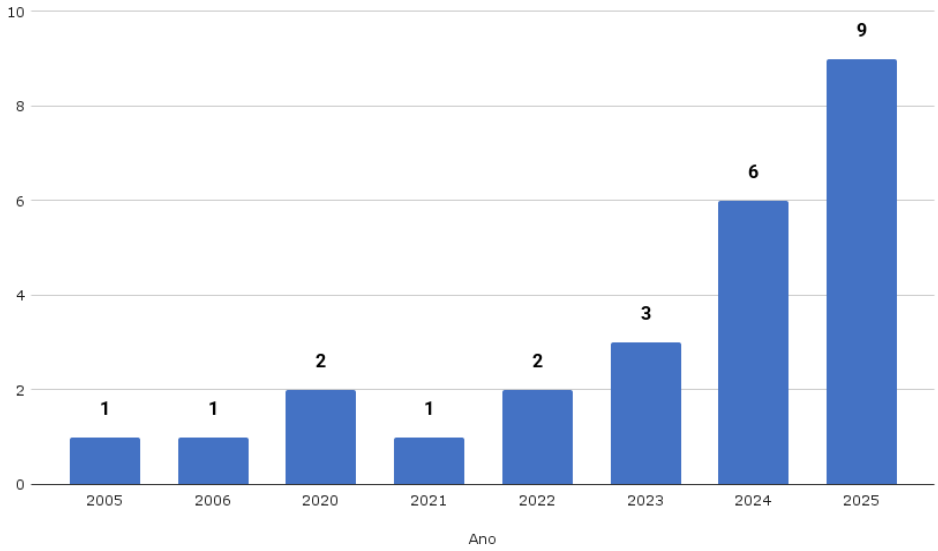


Figura 3 – Etapas da revisão
Fonte: Autores (2026).

No gráfico 1, é possível visualizar a distribuição anual dos artigos sobre a temática no período de 2005 a 2025:

Gráfico 1 – Documentos aceitos por ano de publicação



Fonte: Autores (2026).

A análise da distribuição anual das publicações sobre a temática evidencia um desenvolvimento tardio e assimétrico na literatura científica, indicando um início incipiente e ainda pouco estruturado. Após esse período inicial, nota-se uma lacuna prolongada de publicações até 2020. A partir de 2020, verifica-se uma retomada com crescimento gradual no número de publicações. Esse aumento torna-se mais expressivo a partir de 2023 e se intensifica nos anos de 2024 e 2025, que concentram, respectivamente, seis e nove publicações. Tal tendência pode indicar a consolidação recente do tema, impulsionada pelo crescimento do uso de técnicas de IA, pela demanda por dados sintéticos mais representativos e pela necessidade de incorporar estruturas semânticas na geração desses dados. Assim, a distribuição anual revela que a relação entre dados sintéticos e os campos da Representação da Informação e da Organização do Conhecimento configura-se como um campo emergente e de elevado potencial de expansão e aprofundamento teórico-metodológico.

O Quadro 2 apresenta a lista de documentos aceito, acompanhada de uma breve descrição de seus objetivos:

Quadro 2 – Documentos aceitos para compor o corpus da pesquisa

Autores	Titulo	Descrição

Jeske *et al.* (2005)	Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems	Utiliza dados sintéticos para geração de conjuntos de dados com o intuito de avaliar a precisão de sistemas de descoberta de conhecimento.
Lin *et al.* (2006)	Development of a Synthetic Data Set Generator for Building and Testing Information Discovery Systems	Utiliza grafos semânticos para apoiar a criação de dados sintéticos mais significativos, voltados à detecção de comportamentos anômalos, como fraudes em cartões de crédito.
Feng *et al.* (2020)	A Schema-Driven Synthetic Knowledge Graph Generation Approach with Extended Graph Differential Dependencies (GDDxs)	Propõe a geração de grafos de conhecimento (GCs) sintéticos com o objetivo de reproduzir padrões de grafos reais em domínios com restrições de privacidade.
Linjordet & Balog (2020)	Sanitizing Synthetic Training Data Generation for Question Answering over Knowledge Graphs	Analisa o uso de dados sintéticos para treinamento e avaliação de modelos de perguntas e respostas.
Agarwal *et al.* (2021)	Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training	Apresenta a conversão de um grande GC em texto natural, demonstrando como essa abordagem pode melhorar o desempenho e a qualidade de modelos de linguagem em tarefas de perguntas e respostas.
Organisciak & Ryan (2022)	Improving text relationship modelling with artificial data	Utiliza dados sintéticos para ampliar o corpus de treinamento e compensar a disponibilidade limitada de dados rotulados. Os dados são aplicados para treinamento de uma ferramenta para classificação de relações em bibliotecas digitais.
Platzer & Krchova (2022)	Rule-adhering synthetic data - the lingua franca of learning	Propõe a geração de dados sintéticos com incorporação de conhecimento de domínio.
Kotal *et al.* (2023)	Knowledge infusion in privacy preserving data generation	Estudo aplicado que demonstra a relação entre dados sintéticos e representação do conhecimento por meio do uso de GCs.
Bikeyev (2023)	Synthetic Ontologies: A Hypothesis	Propõe a criação ontologias sintéticas baseados em GC.
Hubert *et al.* (2023)	Pygraft: Configurable generation of synthetic schemas and knowledge graphs at your fingertips	Apresenta uma ferramenta que gera automaticamente esquemas e GCs sintéticos.
Zhang *et al.* (2024)	Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models	Propõe um método de geração de conhecimento sintético para organizar e integrar informações externas em modelos de linguagem.
Schulz *et al.* (2024)	Learning debiased graph representations from the OMOP common data model for synthetic data generation	Propõe um método de geração de dados sintéticos de pacientes a partir de registros reais de prontuários, com preservação da privacidade e possibilidade de intervenção por especialistas.
Chatterjee *et al.* (2024)	Semantic representation and comparative analysis of physical activity sensor observations using MOX2-5 sensor in real and synthetic datasets: a proof-of-concept-study	Propõe o uso de ontologias e dados sintéticos para representar e ampliar conjuntos de dados de atividades físicas coletados por sensores vestíveis.

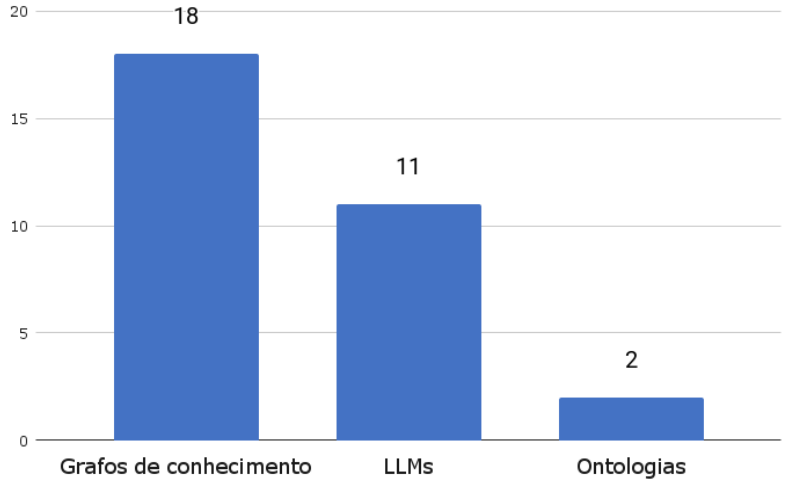
Vuth *et al.* (2024)	KGAST: From Knowledge Graphs to Annotated Synthetic Texts	Propõe uma ferramenta para criação automática de dados sintéticos, utilizando GCs e modelos de linguagem.
Fu *et al.* (2024)	Are Synthetic Time-series Data Really not as Good as Real Data?	Propõe um framework de geração de dados sintéticos para séries temporais com o objetivo de apoiar o treinamento de modelos sem dependência de dados reais.
Kotal *et al.* (2024)	KiNETGAN: Enabling Distributed Network Intrusion Detection through Knowledge-Infused Synthetic Data Generation	Propõe um framework para criação de dados sintéticos de tráfego de rede, com o objetivo de apoiar a detecção de intrusões em sistemas IoT/CPS preservando a privacidade dos dados.
Wang *et al.* (2025)	A Graph-Based Synthetic Data Pipeline for Scaling High-Quality Reasoning Instructions	Propõe um pipeline de geração de dados sintéticos baseado em GCs para ampliar e organizar dados de raciocínio, apoiando o treinamento contínuo de modelos de linguagem.
Chen *et al.* (2025)	GraphGen: Enhancing Supervised Fine-Tuning for LLMs with Knowledge-Driven Synthetic Data Generation	Propõe um framework de geração de dados sintéticos guiado por GCs.
Suzuki *et al.* (2025)	Synthetic Data Generation for Book Recommendation Using Knowledge Graph Embedding	Propõe um método para gerar dados sintéticos para recomendação de livros a partir de GCs.
Ma *et al.* (2025)	Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continue Pre-training of Large Language Models	Propõe um framework de geração de dados sintéticos baseado em grafos de contexto, que explora associações de conhecimento entre documentos.
Marchesin *et al.* (2025)	Large language models and data quality for knowledge graphs	Discute o uso de GCs para criação de dados sintéticos e a importância da avaliação da qualidade de GCs.
Yu *et al.* (2025)	KGSynX: Knowledge Graph and Explainable Feedback Guided LLMs for Synthetic Tabular Data Generation	Propõe um método de geração de dados sintéticos tabulares baseado em GCs.
Fu *et al.* (2025)	Utilizing Language Models for Synthetic Knowledge Graph Generation	Propõe um framework para geração de GCs sintéticos, utilizando modelos de linguagem para ampliar e complementar conjuntos de dados reais de forma estruturada.
Yan (2025)	A Novel Pipeline for Generating Realistic Synthetic CDISC ADaM Datasets Using Large Language Models and Knowledge Graphs	Propõe um pipeline para geração de dados clínicos sintéticos no padrão ADaM usando modelos de linguagem e GCs.
Jing *et al.* (2025)	KG-Enhanced Synthetic Report Generation for Addressing Class Imbalance in Aviation Safety Data	Propõe a geração de dados sintéticos voltado para segurança da aviação, integrando modelos de linguagem e GCs.

Fonte: Autores (2026).

A análise dos estudos aceitos mostrou que a geração de dados sintéticos evoluiu de abordagens genéricas para métodos cada vez mais guiadas por conhecimento, nas quais GCs, ontologias e modelos de linguagem desempenham papel central. O gráfico 2 mostra

a quantidade de artigos aceitos que se utilizam de GCs, LLMs e ontologias na geração de dados sintéticos:

Gráfico 2 – Uso de GCs, LLMs e ontologias nos estudos aceitos



Fonte: Autores (2026).

A análise dos estudos aceitos evidencia que a mera reprodução de propriedades estatísticas dos dados reais não é suficiente para gerar dados sintéticos realistas, representativos e confiáveis. Nesse contexto, a incorporação de estruturas semânticas, relações e regras específicas do domínio mostra-se essencial para a produção de dados sintéticos mais consistentes.

Diante disso, torna-se necessário recorrer a abordagens que permitam estruturar e preservar os significados e relações contextuais adjacentes aos dados, e não apenas suas características quantitativas. Nesse cenário, mecanismos de representação e organização do conhecimento podem assumir papel central, ao possibilitar a modelagem formal de entidades, relações e restrições de domínio, oferecendo suporte conceitual e técnico para a geração de dados sintéticos semanticamente enriquecidos e alinhados aos objetivos das aplicações de IA.

5.1. Representação do conhecimento e Organização do Conhecimento no contexto de dados sintéticos: Grafos de conhecimento em foco

Como observado nos estudos aceitos, a relação entre as áreas discutidas e contexto dos dados sintéticos para treinamento de IA ocorre principalmente no processo de criação de dados sintéticos, baseada na necessidade de dados mais representativos.

Uma das possibilidades é o uso de modelos de IA para a criação desses dados, com a aplicação de LLMs, entretanto, mesmo com a aplicação desses modelos, aplicações em domínios específicos ainda enfrentam desafios significativos (Chen *et al.*, 2025).

Exemplos desses desafios englobam dados de domínios específicos, que exigem altos níveis de qualidade e precisão, onde a quantidade de dados originais não é suficiente para promover treinamento adequado dos modelos ou esbarram em aspectos éticos e questões de privacidade, como nas áreas da saúde e da segurança pública (Chen *et al.*, 2025; Hubert *et al.*, 2024; Feng *et al.*, 2020).

Para lidar com esse cenário, uma das propostas da literatura é aliar os LLMs, e outras estratégias de IA, com estratégias de Representação do Conhecimento. Nesse contexto, o termo representação do conhecimento adquire o significado de:

> Estudo de como as crenças, intenções e julgamentos de um agente inteligente podem ser expressos adequadamente para o raciocínio automatizado. Representação do Conhecimento e Raciocínio (RCR) significa informação do mundo real que um computador pode entender e aplicar para lidar com situações desafiadoras no mundo real. (Pinto, 2022, p. 11)

Um dos resultados desse processo são os grafos "que oferecem um método alternativo com foco em interpretabilidade, verificabilidade e intervenção por meio de especialistas humanos" (Schulz *et al.*, 2024, p. 2).

Um Grafo de Conhecimento (GC) pode ser definido como: "uma representação organizada de entidades do mundo real e seus relacionamentos" (Stegeman, 2024, não paginado, tradução nossa). Esses grafos podem ser armazenados em bancos de grafos, e também podem ser conhecidos como redes semânticas, "pois representam uma rede de entidades do mundo real, como objetos, eventos, situações ou conceitos, e ilustra a relação entre eles" (IBM, 2021). Nesse contexto, os GCs:

> geralmente são compostos de conjuntos de dados de várias fontes, que frequentemente diferem em estrutura. Esquemas, identidades e contexto trabalham juntos para fornecer estrutura a diversos dados. Os esquemas fornecem a estrutura para o gráfico de conhecimento, as identidades classificam os nós subjacentes adequadamente e o contexto determina a configuração na qual esse conhecimento existe. Esses componentes ajudam a distinguir palavras com vários significados. (IBM, 2021)

Os GCs se destacam por permitirem o armazenamento do conjunto de dados juntamente a estrutura que provê a eles significado, por meio de uma estrutura conceitual flexível, que ressalta os relacionamentos existentes e permite a organização do conhecimento de determinado domínio visando aplicação em contextos diversos (Stegeman, 2024).

Esses esquemas ou estruturas podem ser simples ou complexos, baseados ou não na construção de vocabulários semânticos e ontologias. Nesse contexto, destaca-se como desafio terminológico a conceituação dos SOCs, em especial no que tange aos Vocabulários e as Ontologias.

Para o estabelecimento de GCs, os vocabulários exercem o papel de estrutura, sendo entendidos como uma coleção de termos criados para um propósito específico, composto por classes e propriedades que representam um domínio particular (Bizer*et al.*, 2009; W3C, 2017; W3C, 2013).

> Os vocabulários: definem os conceitos e relacionamentos (também chamados de "termos" ou "atributos") usados para descrever e representar uma área de interesse. Eles são usados para classificar os termos que podem ser usados em uma aplicação específica, caracterizar possíveis relacionamentos e definir possíveis restrições ao uso desses termos. (W3C, 2017, não paginado, tradução nossa)

Nesse contexto, torna-se complexa a diferenciação entre diferentes tipos de SOC, em especial com as ontologias. Os termos são muitas vezes adotados pela comunidade como sinônimos, aplicados de maneira intercambiável.

O W3C (2008, não paginado, tradução nossa), aponta que "vocabulário e ontologia são usados indistintamente no contexto desta especificação". O glossário de termos do W3C (2011, não paginado, tradução nossa), ao se referir ao termo vocabulário, menciona que: "O uso deste termo se sobrepõe ao de Ontologia". O W3C (2017, não paginado, tradução nossa) ao abordar a relação entre os termos ontologia e vocabulário afirmam que "não há uma divisão estrita entre os artefatos referidos por esses nomes" e que em relação ao termo vocabulário "o uso deste termo se sobrepõe ao de Ontologia".

Neste sentido, é possível afirmar que:

> Não existe uma diferenciação rígida entre os termos "vocabulário" e "ontologia", sendo comumente utilizado pelo W3C e pela comunidade de maneira intercambiável. Entretanto, é possível considerar o termo vocabulário como um termo abrangente, do qual são tipos ontologias e vocabulários controlados/sistemas de organização do conhecimento. Nessa acepção, as ontologias são vocabulários caracterizados pela complexidade formal, pela presença de axiomas e restrições e pelo caráter estruturante, fornecendo a base para a descrição de outros vocabulários. (Jesus, 2025, p. 124)

Existem diferentes estruturas para representar e armazenar GCs. Para suporte na criação de dados sintéticos, destacam-se os GCs baseados em RDF e os *Property Graphs* (PG):

O RDF é um framework (estrutura), ou modelo padrão de dados, que permite identificar e descrever as características dos recursos na Web, bem como explicitar e

nomear as relações existentes entre recursos. Baseia-se na estrutura de declarações em formato de tripla. O sujeito é representado por um URI do recurso cuja caraterística ou relação está sendo descrita. O predicado deve ser o URI da propriedade de um vocabulário que nomeie a característica ou a relação descrita. O objeto pode ser um URI ou um Literal que represente o valor da propriedade para a característica descrita ou um outro recurso com o qual o primeiro possui uma relação. (Jesus, 2025, p. 120)

Um GC é composto por um conjunto de triplas, que pode ser reunido em um catálogo de dados RDF.

As aplicações em RDF apresentam alguns desafios para criação de dados sintéticos, "Porque o RDF não foi projetado pensando em sistemas de banco de dados. Para começar, não é possível identificar relacionamentos únicos do mesmo tipo entre dois nós em RDF." (Webber, 2024, não paginado, tradução nossa). Uma alternativa são os PG.

Os PGs são estruturados de maneira que a informação:

> É organizada como nós, relacionamentos e propriedades em um grafo de propriedades. Os nós são etiquetados com um ou mais rótulos, identificando seu papel na rede. Os nós também podem armazenar qualquer número de propriedades como pares chave-valor. Os relacionamentos fornecem conexões direcionadas e nomeadas entre dois nós. Os relacionamentos sempre têm uma direção, um tipo, um nó inicial e um nó final, e podem ter propriedades, assim como nós. Embora os relacionamentos sejam sempre direcionados, eles podem ser navegados eficientemente em qualquer direção. (Webber, 2024, não paginado, tradução nossa)

Os PGs também podem adotar ontologias e vocabulários em sua estruturação, destacando assim a importância desses instrumentos para a criação de dados sintéticos mais representativos.

5.2 OI e RI no contexto de dados sintéticos: desafios e questionamentos

Não foram identificadas na literatura recuperada discussões focadas em abordar a relação entre OI e RI e o contexto dos dados sintéticos para treinamento de IA. Entretanto, uma série de questionamentos podem ser levantados, demonstrando a relevância de que a relação entre essas áreas seja discutida e aprofundada.

> O primeiro questionamento a ser levantado é em relação à necessidade de representar, armazenar e recuperar dados sintéticos, que para ser melhor discutido depende de outro questionamento: como ocorre o ciclo de vida dos dados sintéticos. Entender o ciclo de vida permite uma melhor compreensão sobre a natureza dos dados, além de fornecer um vocabulário compartilhado que permite a diferentes profissionais discutirem questões essenciais relacionadas à publicação e consumo de dados na Web. Além disso, um ciclo de vida de dados ajuda a explicar mudanças de paradigma, a comparar a funcionalidade de

> diferentes plataformas e a auxiliar na integração de esforços de implementação previamente díspares. (Santos *et al.*, 2018, p. 2, tradução nossa)

Nas duas etapas da presente pesquisa não foram recuperados ciclos de vida relacionados especificamente a dados sintéticos, que abordem os aspectos relacionados aos processos de criação, armazenamento, recuperação e descarte desses dados.

Com base na análise teórica apresentada, foi possível observar em relação a sua criação, que os dados sintéticos são muito heterogêneos, podendo ser criados com a aplicação de metodologias estatisticamente básicas ou com a aplicação de LLMs e modelos de IA complexos.

A depender da complexidade do domínio, da tarefa a ser desenvolvida e dos níveis de qualidade requeridos, os dados sintéticos podem ou não exigir o uso de sistemas de organização do conhecimento ou na aplicação de GCs. Os dados podem ainda ser ou não derivados de dados reais, existindo aplicações que não se baseiam diretamente em dados reais para o treinamento do modelo usado na criação desses dados.

Observou-se ainda que esses dados normalmente são criados apenas para uma aplicação específica, como o treinamento do modelo, não sendo comum discussões sobre o processo de compartilhamento e reuso desses dados.

Embora inicialmente esse aspecto efêmero dos dados sintéticos possa sugerir que, na maioria dos casos, sua representação para recuperação de dados não seria necessária, compreender todos os aspectos do processo de criação dos dados sintéticos mencionados anteriormente é indispensável, uma vez que eles irão impactar diretamente nos resultados das aplicações nas quais serão utilizados.

Mesmo que ao final de seu ciclo de vida os dados sintéticos possam ser descartados, destaca-se a necessidade de manutenção dos metadados que os representam, para que esses possam ser melhor compreendidos futuramente.

Considerando que a construção de metadados que representem o conteúdo e os aspectos técnicos relacionados a dados sintéticos são relevantes, têm-se ainda o questionamento: como representar adequadamente esse tipo de dados?

Não foram identificados na literatura instrumentos, como padrões de metadados e vocabulários semânticos criados especificamente para representar dados sintéticos. Tem-se, portanto, os questionamentos: instrumentos já estabelecidos podem ser adaptados para a representação desses dados? Quais características específicas de dados sintéticos precisam ser levadas em conta na adaptação/criação de instrumentos para a representação desses dados?

Com base nas discussões e questionamentos apresentados, entende-se que embora não abordada diretamente, a relação entre OI e RI e o contexto dos dados sintéticos é

implícita é necessária, uma vez que esses dados podem ser melhor compreendidos e aplicados quando adequadamente representados.

Além dos aspectos relacionados à representação dos dados sintéticos propriamente ditos, observa-se que não foi explorado pela literatura o uso de metadados e padrões de metadados no processo de criação de dados sintéticos, outro aspecto que pode ser melhor explorado e contribuir para a criação de dados sintéticos mais representativos.

7. Considerações

A presente pesquisa foi conduzida com o objetivo de apresentar e discutir conceitos relacionados aos dados sintéticos e identificar sua potencial relação com as áreas de Organização da Informação, Organização do Conhecimento, Representação da Informação e Representação do Conhecimento.

Na primeira etapa da pesquisa foram discutidos os principais conceitos relacionados a dados sintéticos, apresentada a sua relevância para o treinamento de IAs e seu processo de criação e uso.

Em relação a esse aspecto, observou-se que a criação de dados sintéticos é um processo heterogêneo, tanto em relação aos procedimentos e tecnologias utilizados na sua criação como na necessidade de adoção de instrumentos que ampliem a qualidade e representatividade desses dados. Observou-se ainda que os dados normalmente são criados visando um uso específico, não sendo discutido diretamente na literatura o seu potencial de reutilização.

Na primeira etapa da pesquisa também foram discutidas as áreas de OI, OC, RI e RC, destacando-se os processos e instrumentos relacionados a essas áreas que podem ter impacto no contexto da criação de dados sintéticos para IA, como os SOCs, os metadados e os padrões de metadados.

A segunda etapa da pesquisa concentrou-se em identificar como a relação entre OI, OC, RI e RC os dados sintéticos para o treinamento de IAs têm sido abordados pela literatura.

Observou-se que a discussão dessa relação tem sido abordada de maneira crescente pela comunidade científica, com abordagens interdisciplinares concentradas nas áreas de tecnologia, não tendo sido identificados estudos produzidos no âmbito da Ciência da Informação.

A literatura analisada se concentra principalmente na relação com a RC e OC, abordando o uso de SOC, sendo o principal foco das discussões os GCs.

Essas estruturas são utilizadas nesse contexto visando a criação de dados sintéticos mais representativos e ainda para superar desafios como cenários de escassez de dados para treinamento e de barreiras relacionadas à privacidade dos dados reais.

Nesse sentido, destaca-se como um dos desafios a serem enfrentados na relação entre OC, RC, OI e RI e o contexto dos dados sintéticos para treinamento de IA é a própria barreira terminológica, muitos conceitos semelhantes são trabalhados, como representação e organização do conhecimento, ontologias e vocabulários, mas não possuem as mesmas acepções. Portanto, ressalta-se a importância de um aprofundamento terminológico, para que aproximações possam ser estabelecidas e diferenças possam ser melhor compreendidas, favorecendo assim a cooperação entre as áreas.

Como pesquisas futuras, destaca-se a necessidade de mapeamento e aprofundamento em relação ao ciclo de vida dos dados sintéticos, abordando aspectos como a necessidade de representação e armazenamento desses dados, as possibilidades de reuso e compartilhamento e os aspectos de qualidade de dados a eles relacionados.

Os problemas aqui apresentados se iniciam na discussão da aplicabilidade dos instrumentos de organização e representação da informação ao contexto dos dados, no seu entendimento enquanto objetos informacionais e nas particularidades de sua representação e organização.

Uso de IA generativa

Os autores declaram que não utilizaram ferramentas de inteligência artificial generativa na elaboração deste trabalho.

References

Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2021). Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 3554–3565. http://dx.doi.org/10.18653/v1/2021.naacl-main.278.

Allken, V., Handegard, N. O., Rosen, S., Schreyeck, T., Mahiout, T., & Malde, K. (2018). Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science, 76*(1), 342–349. https://doi.org/10.1093/icesjms/fsy147.

Alves, R. C. V. (2010). Metadados como elementos do processo de catalogação (*Tese de doutorado, Universidade Estadual Paulista*). Repositório Institucional UNESP. http://hdl.handle.net/11449/103361.

Bikeyev, A. (2023). Synthetic ontologies: A hypothesis. *SSRN Electronic Journal*, 1-8. http://dx.doi.org/10.2139/ssrn.4373537.

Boutros, F., Struc, V, Fierrez, J. & Damer, N. (2023). Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing, 135*, e104688. https://dx.doi.org/10.1016/j.imavis.2023.104688.

Brascher, M., & Café, L. (2008). Organização da informação ou organização do conhecimento. *Encontro Nacional de Pesquisa em Ciência da Informação*, *9*, 1-14. http://enancib.ibict.br/index.php/enancib/ixenancib/paper/viewFile/3016/2142.

Chatterjee, A., Prinz, A., Gerdes, M. W., Prinz, A., Riegler, M. A. & Martinez, S. G. (2024). Semantic representation and comparative analysis of physical activity sensor observations using MOX2-5 sensor in real and synthetic datasets: A proof-of-concept-study. *Scientific Reports, 14*(1), e4634. http://dx.doi.org/10.1038/s41598-024-55183-6.

Chen, J., & Little, J. J. (2019). Sports camera calibration via synthetic data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 2497–2504. https://doi.org/10.1109/cvprw.2019.00305.

Chen, Z., Wang, J., & Li, J. (2025). GraphGen: Enhancing supervised fine-tuning for LLMs with knowledge-driven synthetic data generation (*arXiv:2505.20416*). arXiv. http://dx.doi.org/10.48550/ARXIV.2505.20416.

Coneglian, C. S., Santarem Segundo, J. E., & Sant'Ana, R. C. G. (2017). Big Data: Fatores potencialmente discriminatórios em análise de dados. *Em Questão, 23*(1), 62–86. https://doi.org/10.19132/1808-5245231.62-86.

Dahmen, J., & Cook, D. (2019). SynSys: A synthetic data generation system for healthcare applications. *Sensors, 19*(5), e1181. https://doi.org/10.3390/s19051181.

Dwivedi, Y. K., *et al.* (2023). "So what if ChatGPT wrote it?": Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, e102642. http://dx.doi.org/10.1016/j.ijinfomgt.2023.102642.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*(2), 293–314. https://doi.org/10.48550/arXiv.1308.1479.

Feng, Z., Mayer, W., He, K., Kwashie, S., Stumptner, M. & Grossmann, G. (2021). A schema-driven synthetic knowledge graph generation approach with extended graph differential dependencies (GDDxs). *IEEE Access, 9*, 5609–5639. http://dx.doi.org/10.1109/access.2020.3048186.

Fu, F., Chen, J., Zhang, J., Yang, C., Ma, L., & Yang, Y. (2024). Are synthetic time-series data really not as good as real data? (*arXiv:2402.00607*). arXiv. http://dx.doi.org/10.48550/ARXIV.2402.00607.

Fu, S., Mai, P., Jingqi, Z., & Pang, Y. (2025). Utilizing language models for synthetic knowledge graph generation. *ICLR 2025 Workshop on Data Problems*. https://openreview.net/forum?id=IutH9tRtMI.

Gartner. (2024). Os especialistas da Gartner respondem às principais perguntas sobre IA generativa para a sua empresa. https://www.gartner.com.br/pt-br/temas/inteligencia-artificial-generativa.

Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics, 13*(17), e3509. http://dx.doi.org/10.3390/electronics13173509.

Håkansson, A., & Phillips-Wren, G. (2024). Generative AI and large language models - Benefits, drawbacks, future and recommendations. *Procedia Computer Science, 246*, 5458–5468. http://dx.doi.org/10.1016/j.procs.2024.09.689.

Hubert, N., Monnin, P., d'Aquin, M., Monticolo, D., & Brun, A. (2023). PyGraft: Configurable generation of synthetic schemas and knowledge graphs at your fingertips. *European Semantic Web Conference, 1*, 3–20. http://dx.doi.org/10.48550/ARXIV.2309.03685.

IBM. (2021) O que é um gráfico de conhecimento? https://www.ibm.com/br-pt/think/topics/knowledge-graph.

Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. *Big Data & Society, 10*(1), 1–12. https://doi.org/10.1177/20539517221145372.

Jeske, D. R., *et al.* (2005). Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 5*, 756–762. https://doi.org/10.1145/1081870.1081969.

Jesus, A. F. de. (2025). Qualidade de dados Linked Data para seleção de fontes e criação de links: estudo teórico, terminológico e processual. (*Tese de doutorado, Universidade Estadual Paulista*). Repositório Institucional UNESP. https://hdl.handle.net/11449/316194.

Jing, X., Bhanpato, J., Bendarkar, M., V., & Mavris, D. (2025). KG-enhanced synthetic report generation for addressing class imbalance in aviation safety data. *AIAA AVIATION Forum and ASCEND 2025*. http://dx.doi.org/10.2514/6.2025-3250.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic Data—what, why and how? *arXiv*. https://doi.org/10.48550/arXiv.2205.03257.

Kotal, A., Das, N., & Joshi, A. (2023). Knowledge infusion in privacy preserving data generation. *KDD Workshop on Knowledge-infused Learning,* 29TH ACM SIGKDD. https://ebiquity.umbc.edu/paper/html/id/1136/Knowledge-Infusion-in-Privacy-Preserving-Data-Generation.

Kotal, A., Luton, B., & Joshi, A. (2024). KiNETGAN: Enabling distributed network intrusion detection through knowledge-infused synthetic data generation. *2024 IEEE 44th International Conference on Distributed Computing Systems Workshops (ICDCSW), 44*, 140–145. http://dx.doi.org/10.48550/ARXIV.2405.16476.

Lima, G. A. (2020). Organização e representação do conhecimento e da informação na web: teorias e técnicas. *Perspectivas Em Ciencia Da Informacao, 25*, 57-97. https://periodicos.ufmg.br/index.php/pci/article/view/22283.

Lin, P. J., Samadi, B., Cipolone, A., Jeske, D., R., Cox, S., & Rendon, C. (2006). Development of a synthetic data set generator for building and testing information discovery systems. Third *International Conference on Information Technology: New Generations (ITNG'06), 3*, 707–712. http://dx.doi.org/10.1109/itng.2006.51.

Linjordet, T., & Balog, K. (2020). Sanitizing synthetic training data generation for question answering over knowledge graphs. *arXiv, 20*, 1–8. http://dx.doi.org/10.48550/ARXIV.2009.04915.

Ma, S., Jiang, X., Xu, C., Yang, C., Zhang, L., & Guo, J. (2025). Synthesize-on-Graph: Knowledgeable synthetic data generation for continue pre-training of large language models (*arXiv:2505.00979*). arXiv. http://dx.doi.org/10.48550/ARXIV.2505.00979.

Marchesin, S., Silvello, G., & Alonso, O. (2025). Large language models and data quality for knowledge graphs. *Information Processing & Management, 62*(6), e104281. http://dx.doi.org/10.1016/j.ipm.2025.104281.

Moreira, W. (2018). *Sistemas de organização do conhecimento: Aspectos teóricos, conceituais e metodológicos* (Tese de Livre-Docência, Universidade Estadual Paulista). Repositório Institucional UNESP. repositorio.unesp.br.

Nikolenko, S. I. (2021). Synthetic data for deep learning. Springer.

Organisciak, P., & Ryan, M. (2022). Improving text relationship modelling with artificial data. *Journal of Information Science, 50*(2), 434–446. http://dx.doi.org/10.1177/01655515221093031.

Pezoulas, V. C., Zaridis, D. I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N. S., & Fotiadis, D. I. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*, *23*, 2892–2910. https://doi.org/10.1016/j.csbj.2024.07.005.

Platzer, M., & Krchova, I. (2022). Rule-adhering synthetic data -- the lingua franca of learning (*arXiv:2209.06679*). arXiv. http://dx.doi.org/10.48550/ARXIV.2209.06679.

Santos, P. L. V. A. D. C., & Sant'Ana, R. C. G. (2013). Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. *Ciência da Informação, 42*. https://ru.dgb.unam.mx/items/3c0cf793-2ce0-4da5-a992-46efb501c541.

Schulz, N. A., *et al.* (2024). Learning debiased graph representations from the OMOP common data model for synthetic data generation. *BMC Medical Research Methodology, 24*(1), 1–13. http://dx.doi.org/10.1186/s12874-024-02257-8.

Shermeyer, J., Hossler, T., Etten, A., V., Hogan, D., Lewis, R., & Kim, D. (2021). RarePlanes: Synthetic data takes flight. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 5*, 207–217. https://doi.org/10.1109/wacv48630.2021.00025.

Stegeman, J. (2024) What Is a Knowledge Graph? *Neo4j Blog* https://neo4j.com/blog/genai/what-is-knowledge-graph/.

Suzuki, S., Shibata, H., & Takama, Y. (2025). Synthetic data generation for book recommendation using knowledge graph embedding. *Communications in Computer and Information Science, 2414*, 305–318. http://dx.doi.org/10.1007/978-981-96-4589-3_22.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data*. Epoch AI. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

Vuth, N., Sérasset, G., & Schwab, D. (2024). KGAST: From knowledge graphs to annotated synthetic texts. *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), 1*, 43–55. http://dx.doi.org/10.18653/v1/2024.kallm-1.5.

Wang, J., *et al.* (2025). A graph-based synthetic data pipeline for scaling high-quality data. *International Conference on Learning Representations (ICLR 2025)*, 1-20. https://openreview.net/forum?id=CEE9cAQJ10.

Webber, J. (2024). RDF vs. property graphs: Choosing the right approach for implementing a knowledge graph. *Neo4j Blog*. https://neo4j.com/blog/knowledge-graph/rdf-vs-property-graphs-knowledge-graphs/.

Yan, J. (2025). A novel pipeline for generating realistic synthetic CDISC ADaM datasets using large language models and knowledge graphs. *TechRxiv*. 1-15. http://dx.doi.org/10.36227/techrxiv.174234967.75285658/v1.

Yu, K., *et al.* (2025). KGSynX: Knowledge graph and explainable feedback guided LLMs for synthetic tabular data generation. *24th International Semantic Web Conference, 24*, 215–220. https://ceur-ws.org/Vol-4085/paper34.pdf.

Zhang, J., Cui, W., Huang, Y., Das, K., & Kumar, S. (2024). Synthetic knowledge ingestion: Towards knowledge refinement and injection for enhancing large language models (*arXiv:2410.09629*). arXiv. http://dx.doi.org/10.48550/ARXIV.2410.09629.

Zheng, Z., Cai, Y., & Li, Y. (2016). Oversampling method for imbalanced classification. *Computing and Informatics*, *34*(5), 1017–1037. https://www.cai.sk/ojs/index.php/cai/article/view/1277.